

**Part I**

**The Methods**



# 1

## What can the reader expect from this book?

Experiments are conducted. Data are gathered. Researchers are looking for an effect, a change predicted by their musings over a model. At the very least, they want to gauge the *direction* of change: how much evidence is there in the data for a positive effect? More, they want an estimate of the size of the effect.

The statistical evidence for the direction of change is found in a test statistic. But how does one define and measure this ‘statistical evidence’?

In this book we provide a theory for inference in which the word evidence is central and meaningful. We show how to transform test statistics from different studies onto the same calibration scale where it is easier to measure, interpret and combine the evidence in them. Our approach lays the foundation for a meta-analytic theory with *known* weights. Further, it often leads to accurate confidence intervals for standardized effects using smaller sample sizes than would be achieved through standard asymptotic approximations.

The coming chapters are divided into two parts, dealing with methods and theory, respectively. In this chapter we give a taste of things to come. After introducing the calibration scale for evidence, we apply the methods to data from the meta-analytic review literature. Then we discuss standardized effects, sometimes called effect sizes, for two-sample comparisons, and note that each standardized effect is a simple function of a correlation coefficient.

## 1.1 A calibration scale for evidence

### 1.1.1 $T$ -values and $p$ -values

Consider the simple normal model with unknown mean  $\mu$  and standard deviation 1. Given  $n$  observations  $X_1, \dots, X_n$  one rejects the null  $\mu = 0$  in favor of the alternative  $\mu > 0$  if the sample mean  $S = \bar{X}_n$  is ‘large enough’. The test statistic  $S$  is known to contain the evidence required for the test, but the word evidence is rarely defined. In this case we define the *evidence for the alternative* to be the transformed statistic  $T = \sqrt{n} S = \sqrt{n} \bar{X}_n$ . This  $T$  is normally distributed with mean  $\sqrt{n} \mu$  and standard deviation 1, so  $T$  is an unbiased estimator of its mean  $\sqrt{n} \mu$  with standard error 1.

Note that the expected evidence  $\sqrt{n} \mu$  grows linearly with  $\mu$ , and we require that any definition of evidence for  $\mu > 0$  would grow with  $\mu$ . In addition, the expected evidence grows with the square root of the sample size; this is consistent with the notion from estimation that evidence for an unknown  $\mu$  grows only at this rate: one needs four times as many observations to estimate  $\mu$  with twice the accuracy, because the standard error of  $\bar{X}_n$  is  $1/\sqrt{n}$ .

Thus evidence for the alternative as defined here is a random quantity which always has inherent error, in fact a standard normal error, whether or not the null hypothesis holds. If one observes  $T = 1.645$  and reports this as evidence for the alternative, one should also note the standard error is 1; it is better to write  $1.645 \pm 1$ . When one does this, one realizes that what is sometimes called a ‘significant’ outcome could quite easily have been something else.

Now suppose that one has two independent experiments similar to the one above, with respective sample means  $\bar{X}_1$  based on  $n_1$  observations and  $\bar{X}_2$  based on  $n_2$  observations. How can we combine the evidence in  $T_1 = \sqrt{n_1} \bar{X}_1$  and  $T_2 = \sqrt{n_2} \bar{X}_2$  to obtain a single evidence  $T$  for the alternative  $\mu > 0$ ? A good choice is  $T_{\text{comb}} = (\sqrt{n_1} T_1 + \sqrt{n_2} T_2) / \sqrt{n_1 + n_2}$ , because it is the mean of all  $n_1 + n_2$  observations, rescaled to have variance 1. Also,  $T$  is a linear combination of independent normal variables and hence normal, with expected evidence  $\sqrt{n_1 + n_2} \mu$  and standard deviation 1. It is on the same calibration scale as  $T_1$  and  $T_2$ . In particular if  $n_1 = 9$ ,  $n_2 = 16$  and  $T_1 = 1.645$ ,  $T_2 = 2.236$ , then the combined evidence for  $\mu > 0$  is  $T_{\text{comb}} = 2.848 \pm 1$ .

Another way of combining the evidence for  $\mu > 0$  is to take  $(T_1 + T_2) / \sqrt{2}$  which is normal with mean  $(\sqrt{n_1} + \sqrt{n_2}) \mu / \sqrt{2}$  and variance 1. For the example in which  $T_1 = 1.645$  and  $T_2 = 2.236$  this combination yields  $2.808 \pm 1$ , which is slightly smaller than  $T_{\text{comb}}$ . Note that  $\sqrt{n_1 + n_2}$  is always greater than or equal to  $(\sqrt{n_1} + \sqrt{n_2}) / \sqrt{2}$  and equality is achieved only when  $n_1 = n_2$ . Thus, the first combination of the evidence is on average always at least as good as the second one. The proof of the cited inequality is left to the reader; it follows from the concavity of the square root function.

### Traditional ‘significance’ is only weak evidence for the alternative

So far we only have transformed the test statistic  $S$  onto a scale whose unit equals the standard deviation of  $T = T(S)$ . A traditional marker on this scale is 1.645, the point

dividing ‘significant’ from ‘nonsignificant’ values. But of course there is almost no difference between the results  $T = 1.644$  and  $T = 1.646$ , and adding and subtracting the true standard error of 1 puts evidence into its proper perspective: it has a standard normal error. The result  $T = 1.645 \pm 1$  illustrates that  $T = 1.645$  is unreliable. If forced to give an adjective describing such evidence, we would call it ‘weak’. Twice as much evidence,  $T = 3.3$ , will then be called ‘moderate’, and three times as much evidence,  $T = 5$ , will be called ‘strong’. See Figure 1.1 for plots of some evidence possibilities. These somewhat arbitrary descriptions are necessarily vague because evidence is a random quantity. But we think they are a more realistic guide than setting degrees of ‘significance’ based on  $p$ -values.

The  $p$ -value of an observed value of a test statistic is often thought to be a measure of evidence against a null hypothesis, with smaller values indicating larger evidence. In a certain sense this is true, but the  $p$ -value is conditional on the data from a particular experiment, and so has relevance only for that particular experiment. If one wants to compare  $p$ -values from different experiments, or even to combine the evidence in them as in meta analysis, one must take into account their distributional properties.

First assume the null hypothesis holds. Then the  $p$ -value, when considered as a random variable, is known to have a uniform distribution on the unit interval when the test statistic has a continuous distribution, and nearly uniform if the test statistic has a discrete distribution. So, one might argue, one can indeed combine  $p$ -values using

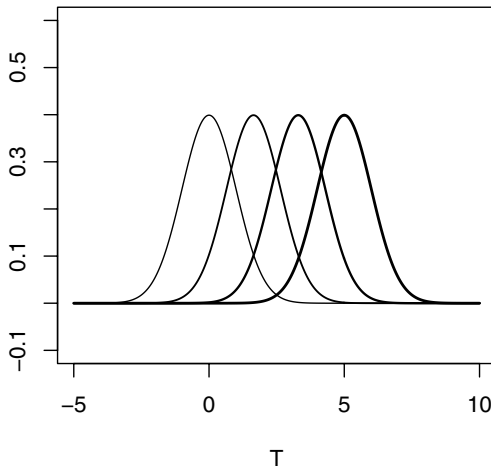


Figure 1.1 The distribution of evidence on the proposed calibration scale is always normally distributed with variance 1. When  $\sqrt{n}\mu = 0$ , the evidence  $T$  is centered on the origin; this is often called the null distribution of  $T$ . Other possibilities are centered on  $\sqrt{n}\mu = 1.645, 3.3$  and  $5$ , respectively shown from left to right. The point is that evidence is a random quantity with an unknown mean but standard normal error. Upon observing  $T = 3.3$ , one should report  $T = 3.3 \pm 1$ . This gives a clear indication not only of the magnitude, but also the error inherent in evidence  $T$ .

their common null uniform distribution, and assumed independence of experiments. But when one has in hand a number of small  $p$ -values, each of which is considered ‘significant’, the conviction grows that the null distribution is indeed false, and what is really desired is a combination of evidence that works whether or not the null hypothesis is true. Such a combination cannot be based on the assumption that the null hypothesis is true and that the  $p$ -value has a rectangular density. These considerations and others, explained in detail in Chapter 16, lead us to the conclusion that  $p$ -values, when considered as random variables, are on the wrong scale for calibration and interpretation of statistical evidence, and for forming a combined conclusion from a set of tests.

Before leaving this section we point out that a  $p$ -value for a test based on the  $T$ -statistic can be obtained if desired through the probit transformation of an observed value  $t$  of the evidence  $T$ . It is  $p = 1 - \Phi(t) = \Phi(-t)$ . For this simple example the  $p$ -values based on  $T = T(S)$  are exactly the same as those based on  $S$ .

### 1.1.2 How generally applicable is the calibration scale?

So far we have only considered the simplest model of testing for a normal mean when the standard deviation is known. The transformation of the test statistic  $S = \bar{X}_n$  to evidence  $T = \sqrt{n} S$  only required multiplication by the square root of the sample size. In general one tries to select a transformation  $h$  of the test statistic  $S$  so that  $T = h(S)$  is on this same unit normal calibration scale. In most routine problems of statistics this goal cannot be achieved completely, but it can be achieved approximately to a surprising degree for one- and two-sample binomial and Poisson models, for one- and two-sample  $t$ -tests and for chi-squared and  $F$ -tests. The first step then is to find the variance stabilizing transformation  $h(S)$  for the particular model of interest, and the results of our and others’ endeavors are presented in coming chapters.

In most cases the resulting evidence  $T$  is approximately normal with standard deviation 1 and mean which can be approximated  $E[T] \doteq \sqrt{n} \mathcal{K}(\delta)$ . Here again  $n$  is the sample size,  $\delta$  is a standardized effect and  $\mathcal{K}$  is the Key Inferential Function. Knowing the Key is like knowing the power function in traditional Neyman–Pearson testing; it contains all the important information about the relationship between the standardized effect  $\delta$  and its transformed value  $\kappa = \mathcal{K}(\delta)$ . This information can be exploited to choose sample sizes to obtain desired amounts of evidence, up to standard error 1, or to derive confidence intervals for  $\delta$ .

#### Example 1. The one-sample $t$ -test

Take  $X_1, \dots, X_n$  independent, each having the normal distribution with unknown mean  $\mu$  and variance  $\sigma^2$ . The raw effect is  $\mu - \mu_0$ , where  $\mu_0$  is a known value determined by scientific context. The standardized effect is  $\delta = (\mu - \mu_0)/\sigma$ . For testing the null  $\mu = \mu_0$  against the alternative  $\mu > \mu_0$  the test statistic  $S = \sqrt{n}(\bar{X}_n - \mu_0)/s_n$  is known to have a Student  $t$ -distribution with  $n - 1$  degrees of freedom under the null hypothesis and a noncentral  $t$  distribution with the same  $n - 1$  degrees of freedom and noncentrality parameter  $\sqrt{n} \delta$ . Chapter 20 contains further details, where

it is also shown that a variance stabilizing transformation  $T = h(S)$  has the property that, to a useful approximation,  $T$  has the  $N(\sqrt{n} \mathcal{K}(\delta), 1)$  distribution for a wide range of values of  $n$  and  $\delta$ .

The Key Inferential Function for this measure of evidence is

$$\begin{aligned} \mathcal{K}(\delta) &= \sqrt{2} \sinh^{-1}(\delta/\sqrt{2}) \\ &= \sqrt{2} \ln(\delta/\sqrt{2} + \sqrt{1 + \delta^2/2}). \end{aligned}$$

This simple monotonic function together with the sample size  $n$  provide all the information required for inference regarding  $\delta$ , provided  $n$  is not too small. For example, when  $n = 10$  accurate 95 % confidence intervals can be derived for any  $\delta$  satisfying  $-2 < \delta < 2$ .

### Example 2. The one-sample binomial test

Let  $X$  have the binomial distribution with parameters  $n, p$  where  $n$  is known and  $0 < p < 1$ . For testing the null  $p = p_0$  against the alternative  $p > p_0$  it is customary to reject the null when the test statistic  $X$  is too large; or equivalently when  $\hat{p} = X/n$  is too large. It is well known (see Chapter 18) that a classic transformation of  $\hat{p}$  to the unit normal calibration scale is given by  $T = h(\hat{p}) = 2\sqrt{n} \{\arcsin(\sqrt{\hat{p}}) - \arcsin(\sqrt{p_0})\}$ , and this transformation is improved if  $\hat{p}$  is replaced by  $\tilde{p} = (X + 3/8)/(n + 3/4)$ . The Key Inferential Function for this transformation is

$$\mathcal{K}(p) = 2\{\arcsin(\sqrt{p}) - \arcsin(\sqrt{p_0})\}.$$

This Key could have been expressed as a function of the raw effect  $p - p_0$  or the standardized effect  $\delta = \sqrt{n} (p - p_0) / \sqrt{p(1-p)}$  because these effects are monotonic functions of  $p$ , but for this example it would be an unnecessary notational complication. In Section 1.2 we illustrate how this arcsine transformation to the calibration scale can be employed to find and combine the evidence in several studies. But first we need to discuss several issues arising when considering more than one study on the same subject.

### 1.1.3 Combining evidence

Return to the simple normal model of Section 1.1.1, where we tacitly assumed that the true effect  $\mu$  was the same for the two studies, instead of the more realistic assumption that  $T_1 \sim N(\sqrt{n_1} \mu_1, 1)$ ,  $T_2 \sim N(\sqrt{n_2} \mu_2, 1)$  where both  $\mu_1, \mu_2$  are unknown. The joint null hypothesis is now  $\mu_1 = 0 = \mu_2$ , and there are many possible alternatives, each possibly requiring a different combination of evidence. For example, the alternative  $\mu_w = (w_1 \mu_1 + w_2 \mu_2) / (w_1 + w_2) > 0$ , for known positive weights  $w_1, w_2$ , suggests a combination  $T_w = c(w_1 T_1 + w_2 T_2)$ , with constant  $c$  chosen so that  $T_w$  has variance 1. And the joint alternative  $\mu_1 > 0$  and  $\mu_2 > 0$  suggests a combination of the form  $T_{\min} = h(\min\{T_1, T_2\})$  where  $h$  is a transformation to the unit normal calibration scale. The best combination for each alternative is a challenging problem in itself, which we do not pursue here. Rather we test or estimate an overall effect.

In traditional meta analysis it is common to assume the  $\mu_k$  values are equal (the fixed effects model); or to assume that the  $\mu_k$  values themselves are a random sample from a  $N(\mu, \gamma^2)$  model (the random effects model), where  $\gamma^2$  is a variance component introduced to explain the variability in  $\mu_k$ . The advantage of these two models is that there is only one parameter of interest  $\mu$ , the overall effect, and one can test hypotheses regarding  $\mu$  or estimate  $\mu$  without all the complications raised in the previous paragraph for fixed unequal effects.

More generally, we have  $K$  independent studies resulting in evidences  $T_k$  which are approximately normal with variance near 1 and  $E[T_k] \doteq \sqrt{n_k} \mathcal{K}(\delta_k)$  for  $k = 1, \dots, K$ . Here  $T_k$  is the evidence for  $\delta_k > 0$  based on  $n_k$  observations in the  $k$ th study, obtained by a suitable variance stabilizing transformation, and  $\mathcal{K}$  is the associated monotonically increasing Key Inferential Function. There is a one-to-one correspondence between each  $\delta_k$  and  $\kappa_k = \mathcal{K}(\delta_k)$ . The *fixed standardized effects model* in which all  $\delta_k = \delta$  is easiest to deal with, because there is only one  $\delta$ , hence one  $\kappa = \mathcal{K}(\delta)$ . One can find the evidence  $T_{\text{comb}} = \sum_k \sqrt{n_k} T_k / \sqrt{N}$ , where  $N = \sum_k n_k$ , as evidence for the alternative  $\kappa > 0$ , and hence also for  $\delta > 0$ . Note that  $T_{\text{comb}} \sim N(\sqrt{N} \mathcal{K}(\delta), 1)$ . One can also use  $T_{\text{comb}} \pm z_{1-\alpha/2}$  to obtain a  $100(1 - \alpha)\%$  confidence interval  $[L, U]$  for  $\kappa$ , and by back-transformation for  $\delta$ , namely  $[\mathcal{K}^{-1}(L/\sqrt{N}), \mathcal{K}^{-1}(U/\sqrt{N})]$ .

In many problems the assumption that all  $\delta_k = \delta$  is untenable, and testable using Cochran's  $Q$  test of homogeneity. In Chapter 24 a variant of Cochran's  $Q$  called  $Q^*$  is applied to the  $\hat{\kappa}_k$ 's to find the evidence  $T_{Q^*}$  for heterogeneity of the  $\kappa_k$ 's and hence the  $\delta_k$ 's. On the basis of this evidence, the researcher may well prefer the following model.

The *random transformed (standardized) effects model* assumes that the  $\kappa_k$ 's are a random sample of size  $K$  from the normal model with mean  $\kappa$  and variance  $\gamma^2$ , with both parameters unknown. Then the conditional distribution of each  $\hat{\kappa}_k$ , given  $\kappa_k$ , is  $N(\kappa_k, 1/n_k)$ , and unconditionally it is  $N(\kappa, \gamma^2 + 1/n_k)$ . Now when the  $n_k$ 's are all equal, or when their reciprocals are negligible compared to  $\gamma^2$ , the  $\hat{\kappa}_k$ 's are just a sample of size  $K$  from a normal population with mean  $\kappa$  and common variance. Let  $\bar{\kappa}$  and  $s_\kappa^2$  denote the sample mean and variance of these transformed standardized effects. The usual  $t$ -test rejects the null  $\kappa = 0$  in favor of  $\kappa > 0$  when the statistic  $S = \sqrt{K} (\bar{\kappa} - 0)/s_\kappa$  is large. The evidence in this statistic for  $\kappa > 0$ , and hence  $\delta > 0$ , is essentially  $T = \sqrt{2K} \sinh^{-1}(S/\sqrt{2K})$ , as shown in Chapter 20.

If one desires to compute a confidence interval for  $\delta$ , one can find a  $t$ -interval  $[L, U]$  for  $\kappa$  first, namely  $\bar{\kappa} \pm t_{K-1, 1-\alpha/2} s_\kappa/\sqrt{K}$ , and then  $[\mathcal{K}^{-1}(L), \mathcal{K}^{-1}(U)]$  for  $\delta$  by back-transformation.

## 1.2 The efficacy of glass ionomer versus resin sealants for prevention of caries

### 1.2.1 The data

The review by Ahovuo-Saloranta *et al.* (2004) contains three studies in which matching molar teeth in the same children formed the basis for paired comparisons. Two

Table 1.1 Summary of three studies by the authors shown. Note the evidence is in conflict, but this should not preclude an analysis; further studies may demonstrate that one sealant is superior to another. References to these three studies and more background can be found in Ahovuo-Saloranta *et al.* (2004).

		Resin sealant					
		+	-	+	-	+	-
Glass Ionomer Sealant	+	378	28	156	6	191	2
	-	3	3	37	7	9	1
		Arrow (1995)		Poulsen (2001a)		Poulsen (2001b)	

types of sealants were applied at random to the pair, and then the teeth were assessed after 24- to 44-month intervals to detect the presence ‘-’ of one or more caries or ‘+’ no caries. The results of these three studies are summarized in Table 1.1.

The *discordant pairs* are those for which the treatment and control responses differ; let  $f$  be the number of (+, -) pairs and  $g$  be the number of (-, +) pairs. In the first study there are  $f = 28$  pairs for which the response was (+, -): there were no caries in one tooth after glass ionomer treatment, while the corresponding tooth receiving resin sealant did have caries. There were  $g = 3$  pairs in which the two treatments led to the opposite results (-, +). The conditional distribution of  $f$ , given  $f + g$  is binomial with parameters  $f + g$  and  $p$ , where  $p$  is the probability that a discordant pair is (+, -). A test of symmetry in treatment control outcomes is a test of  $p = 0.5$ , with alternative  $p > 0.5$  corresponding to the treatment (in this case glass ionomer) having greater probability of ‘+’ within a discordant pair. (See Lachin (2000), p. 180, for more details.) We can now compute the evidence for  $p > 0.5$  in each of the three studies using  $T = 2\sqrt{n} \{ \arcsin(\sqrt{\tilde{p}}) - \arcsin(\sqrt{0.5}) \}$ , where  $\tilde{p} = (X + 3/8)/(n + 3/4)$ .

## 1.2.2 Analysis for individual studies

### 1.2.2.1 Evidence for $p_k > 0.5$ in individual studies

In the first experiment, there are 31 discordant pairs, so conditionally,  $X_1$  has the binomial(31,  $p_1$ ) distribution, where  $p_1$  is the probability that glass ionomer is more effective than the resin sealant in preventing caries in the first experiment. The evidence against  $p_1 = 0.5$  in favor of  $p_1 > 0.5$  is  $T_1 = 5.05$ , displayed in column 3 of Table 1.2; this is what we would call ‘strong’ evidence.

In the second study, the distribution of  $X_2$  is binomial(43,  $p_2$ ), where again,  $p_2$  is the probability that glass ionomer is more effective in this study. The evidence against  $p_2 = 0.5$  in favor of  $p_2 > 0.5$  is  $T_2 = -5.16$ ; that is, the evidence is even stronger than in the first study, but this time in the opposing direction.

Table 1.2 Summary of synthesis of evidence for the sealant data in Table 1.1.

$k$	$X_k$	$n_k$	$\tilde{p}_k$	$T_k$	$T_{1:k}$	$L_k$	$U_k$	$\hat{\kappa}_k$
1	28	31	0.903	5.05	5.05	0.763	0.976	0.548
2	6	43	0.140	-5.16	-0.67	0.057	0.265	-0.560
3	2	11	0.182	-2.12	-1.39	0.029	0.476	-0.230

For the third study, the evidence against  $p_3 = 0.5$  in favor of  $p_3 > 0.5$  is  $T_3 = -2.12$ , which is weak evidence for the alternative  $p_3 < 0.5$ . It is important to remember that all these evidence values have standard error 1.

**Confidence intervals for  $p_k$  in individual studies**

Confidence intervals  $[L_k, U_k]$  for  $p_k$  are based on Equation (18.2), and for confidence 95% are shown in columns 7 and 8 of Table 1.2. Note that they are not centered on  $\tilde{p}_k$ , but are more reliable than intervals based on the standard asymptotic theory of adding and subtracting 1.96 standard errors to  $\hat{p}$ . For more details, see Chapter 18. These intervals suggest that the  $p_k$  are not equal, but nevertheless for completeness we assume this to be the case in the next section.

**1.2.3 Combining the evidence: fixed effects model**

If we were to assume that all  $p_k = p$ , then we could readily combine the evidence in the individual studies for  $p > 0.5$ . The results in column 6 of Table 1.2 are obtained sequentially: the entry in row  $k$  is based on the first  $k$  studies. The first two studies have strong conflicting evidence, and this is reflected by the combined evidence  $T_{12} = (\sqrt{31} T_1 + \sqrt{43} T_2)/\sqrt{74} = -0.67$ , shown in column 6. It is almost negligible. For the three studies, the combined evidence is  $T_{1:3} = (\sqrt{74} T_{12} + \sqrt{11} T_3)/\sqrt{85} = -1.39$ , which is quite weak evidence in favor of the resin sealant. Thus combining evidence on the calibration scale allows for cancelation of conflicting evidence, leading to the correct conclusion that there is no evidence for a common  $p > 0$ .

One can also obtain a confidence interval for  $p$  based on all three studies. Starting with the combined evidence  $T_{1:3} = -1.39$ , a 95 % confidence interval for the expected evidence  $\sqrt{85} \mathcal{K}(p)$  is  $-1.39 \pm 1.96$ , or  $[L, U] = [-3.35, 0.57]$ . Here the key is  $\mathcal{K}(p) = 2\{\arcsin(\sqrt{p}) - \arcsin(\sqrt{0.5})\}$ , so  $\mathcal{K}^{-1}(y) = \sin^2(y/2 + \pi/4)$ . This leads to the 95 % interval  $[\mathcal{K}^{-1}(L/\sqrt{85}), \mathcal{K}^{-1}(U/\sqrt{85})] = [0.32, 0.53]$  for  $p$ .

**1.2.4 Combining the evidence: random effects model**

The transformed effects  $\hat{\kappa}_k = \mathcal{K}(\tilde{p}_k)$  are shown in Table 1.2, and their respective approximate normal  $N(\kappa_k, 1/n_k)$  distributions depicted in Figure 1.2. The sample mean and standard deviation are  $\bar{\kappa} = -0.081$  and  $s_\kappa = 0.569$ . A test for heterogeneity of these transformed effects based on Cochran’s  $Q$  is described in Chapter 24, and

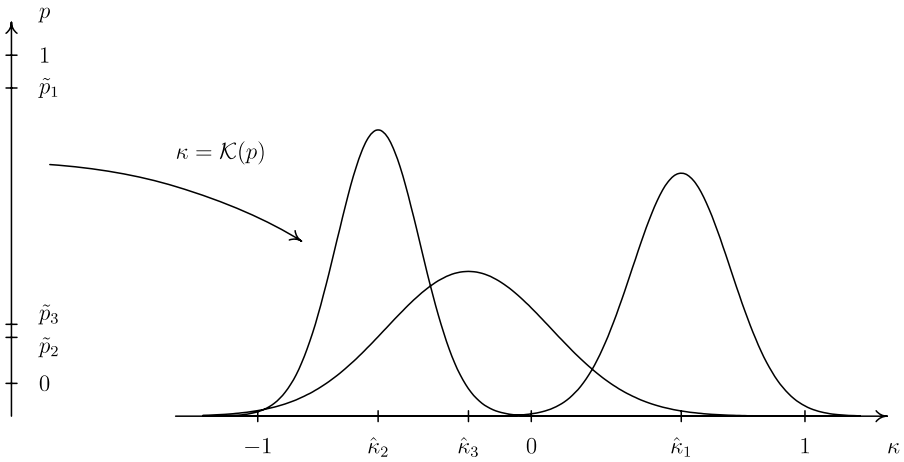


Figure 1.2 Transformation of the estimated probabilities that glass ionomer outperforms resin sealant into transformed effects  $\hat{\kappa}_k = \mathcal{K}(\tilde{p}_k)$ . The evidence  $T_k$  for a positive effect  $p_k - 0.5 > 0$  has distribution that is approximately  $N(\sqrt{n_k} \kappa_k, 1)$ , so  $\hat{\kappa}_k = T_k / \sqrt{n_k}$  has distribution that is approximately  $N(\kappa_k, 1/n_k)$ . These normal distributions are centered at respective unknowns  $\kappa_k$ 's, and depicted in the plot centered at the respective estimates  $\hat{\kappa}_k$ 's.

the evidence for heterogeneity is strong ( $T_{Q^*} \approx 4.5$ ) so a random transformed effects model is in order; it essentially adds a variance component to the model to account for the variability from study to study. Details are given in Section 25.3, where it is shown that if the reciprocals of the sample sizes are small compared to this component, then, even for a small number of studies  $K$ , the evidence for the overall  $\kappa > 0$ , and hence  $p > 0.5$ , is  $T = \sqrt{2K} \sinh^{-1}(S/\sqrt{2K})$ , where  $S = \sqrt{K} (\bar{\kappa} - 0)/s_\kappa$ .

For our data  $S = -\sqrt{3} (0.081/0.569) = -0.25$  and so the evidence  $T$  for  $\kappa > 0$ , and hence  $p > 0$ , is negligible. (Note that here  $T \approx S = -0.25$ , because the function  $\sinh^{-1}$  behaves like the identity near the origin.)

A confidence interval for a representative  $p$  can also be found, starting with the  $t$ -interval for  $\kappa$  of  $\bar{\kappa} \pm t_{2,0.975} s_\kappa / \sqrt{3}$  or  $[L, U] = [-1.49, 1.33]$ . By transforming this interval back via  $\mathcal{K}^{-1}(y) = \sin^2(y/2 + \pi/4)$ , the 95% confidence interval for  $p$  is  $[0.002, 0.986]$ . This interval tells us virtually nothing about  $p$ , but of course this is because the number of studies is small, and the results are contradictory. It confirms that the very strong assumption of a fixed effects model which led to the interval  $[0.32, 0.53]$  for  $p$  is unwarranted.

### 1.3 Measures of effect size for two populations

For us an *effect size* is another term for *standardized effect*; that is, an effect divided by a suitable measure of scale. For a single population with mean  $\mu$ , standard deviation  $\sigma$ ,

it is often taken to be the raw effect  $\mu - \mu_0$  divided by  $\sigma$ . Here  $\mu_0$  is a hypothesized value of  $\mu$  suggested by scientific context. The advantage of standardized effects over raw effects is that they are free of the units of measurement. For two populations with different variances  $\sigma_1^2, \sigma_2^2$ , the question arises of how to standardize the difference of their means  $\Delta = \mu_1 - \mu_2$ . The purpose of this section is to define a standardized effect  $\delta$  for comparing two populations and its associated correlation effect size  $\rho = \rho(\delta)$ .

Let  $X_1, \dots, X_{n_1}$  be a sample of size  $n_1$  from the first population and estimate  $\mu_1$  by the sample mean  $\bar{X}$ ; similarly let  $\bar{Y}$  be based on an independent sample  $Y_1, \dots, Y_{n_2}$  from the second population. Then an unbiased estimator of the effect  $\Delta = \mu_1 - \mu_2$  is  $\hat{\Delta} = \bar{X} - \bar{Y}$ . Now, because  $\hat{\Delta}$  is unbiased for  $\Delta$ , the standard error  $SE[\hat{\Delta}]$  of  $\hat{\Delta}$  satisfies

$$\{SE[\hat{\Delta}]\}^2 = \text{Var}[\hat{\Delta}] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

**Definition 1.1** Let  $N = n_1 + n_2$ , and define the standardized effect by

$$\delta = \frac{\Delta}{\sqrt{N} SE[\hat{\Delta}]}.$$

This effect size  $\delta$  is free of the units of measurement. Note that  $\delta$  is also free of the sample sizes, but does depend on the relative sample sizes, as well as  $\Delta$  and unknown population variances.

There are numerous other definitions of effect sizes in the meta-analytic literature, including those that are Pearson product moment correlations between the variable of interest and a classification variable; this group includes the point-biserial correlation coefficient, see Cohen (1988) and Rosnow and Rosenthal (1996) and references therein. These measures of effect size are often called *correlation effect sizes* and will be denoted generically here by  $\rho$ . Each is related to a corresponding standardized effect  $\delta$  by:

$$\rho = \frac{\delta}{\{1 + \delta^2\}^{1/2}}. \quad (1.1)$$

A plot of  $\rho$  against  $\delta$  is shown in Figure 1.3. Note that  $\rho = \rho(\delta)$  is a strictly increasing function of  $\delta$  with inverse function  $\delta = \delta(\rho) = \rho/\sqrt{1 - \rho^2}$ . In addition,  $\rho$  is an odd function of  $\delta$ ; that is  $\rho(-\delta) = -\rho(\delta)$  for all  $\delta$ .

## Examples

The above Definition 1.1 of standardized effect is employed directly in comparing two normal populations in Chapter 21. Another special case, comparing two Bernoulli populations, is also of interest, and discussed in Chapter 19. Here we reexpress the above results in a simpler notation for this problem. Assume each  $X_i = 1$  or 0, respectively, with probabilities  $p_1, 1 - p_1$ ; that is,  $X_i$  has the Bernoulli( $p_1$ ) distribution, and  $\mu_1 = E[X_i] = p_1$  and  $\sigma_1^2 = p_1(1 - p_1)$ . Similarly let each  $Y_i \sim \text{Bernoulli}(p_2)$ . Then  $\hat{p}_1 = \bar{X}$ ,  $\hat{p}_2 = \bar{Y}$ .

In this context  $\Delta = p_1 - p_2$  and  $\hat{\Delta} = \hat{p}_1 - \hat{p}_2$ . Further, letting  $q = n_2/N$ , where  $N = n_1 + n_2$ , and following the notation of Brown and Li (2005), let  $p = qp_1 +$

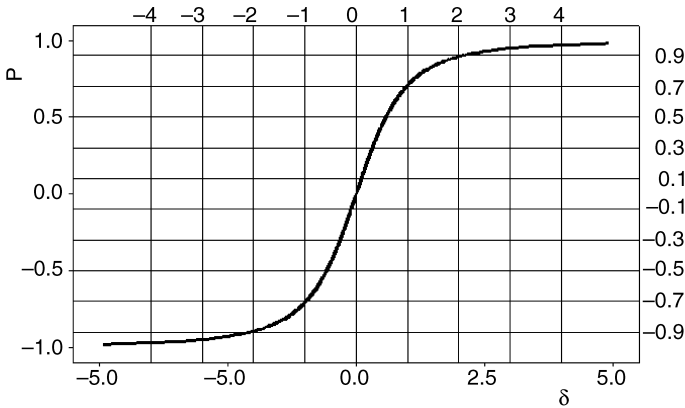


Figure 1.3 The graph of correlation effect size  $\rho$  against standardized effect  $\delta$ .

$(1 - q)p_2$ . They observe that  $N\text{Var}[\hat{\Delta}] = \zeta - \Delta^2$ , where  $\zeta = p(1 - p)/\{q(1 - q)\}$ . The standardized effect is therefore  $\delta = \Delta/\sqrt{\zeta - \Delta^2}$ , and the associated correlation effect size is  $\rho = \Delta/\sqrt{\zeta}$ . The importance of this result to the theory presented here is that in Chapter 19 we define a new and effective variance stabilizing transformation for the risk difference  $\Delta = p_1 - p_2$  and its associated Key Inferential Function is simply  $\mathcal{K}(\rho) = \arcsin(\rho)$ .

### 1.4 Summary

In this text we provide a unified theory of statistical inference in which the word ‘evidence’ is central and meaningful. It grows out of our conviction that the traditional ways of measuring evidence, in particular with probabilities, are neither intuitive nor useful when it comes to making comparisons between experimental results, or when combining them.

We measure evidence for an alternative hypothesis, not evidence against a null. To do this, we have in a sense adopted standardized scores for the calibration scale. Evidence for us is simply a transformation of a test statistic to another one (called *evidence*) whose distribution is close to normal with variance 1, and whose mean grows from 0 with the parameter as it moves away from the null. The transformation required depends on the model, and there is a rich legacy to draw upon from research in the last century.

The advantages of such a theory are many:

- **Conceptual simplicity.** Evidence  $T$  for an alternative is normally distributed with unknown mean and variance 1; it is an unbiased estimator of its mean that always has a standard normal error.

- **Usefulness.** The expected evidence often has the form  $E[T] \doteq \sqrt{n} \mathcal{K}(\delta)$ , where  $\mathcal{K}$  is a known Key Inferential Function. This formula facilitates finding sample sizes required to achieve desired amounts of evidence for an alternative, and deriving confidence intervals for  $\delta$ .
- **Effectiveness.** Compared to methods based on standard asymptotics, these methods generally require smaller sample sizes to achieve good approximations (see Chapter 27).
- **Meta-analytic potential.** Combining evidence on this calibration scale is simpler, because it forms combinations of evidence with *known* weights.

Of course there are disadvantages, too, of which the reader is no doubt aware. One needs to become familiar with square root, arcsine and hyperbolic arcsine transformations. But in this opening chapter we have tried to convey the above listed potential benefits of defining evidence on the unit normal scale. We have sketched the ideas for the most important binomial and normal models, and illustrated the meta-analytic ideas on data from the recent review literature. We have concluded with some relations between effect sizes useful to us in comparing two populations.