Correct interpretation of a p-value in a null hypothesis t-test

One of the central tools used in statistical analysis is hypothesis testing with p-values. For the purposes of this briefing, we are going to use a simple *t-test of the difference in means between two groups* to demonstrate how to interpret a p-value. A random sample of 60 children of a particular age were taken from each of two schools in London, the amount of pocket money they received in a particular week was measured and the following sample statistics were calculated (see Table 1):

Table 1: Pocket money sample statistics

Sample statistics	Group	
	School 1	School 2
Sample mean	£4.88	£5.40
Sample standard deviation	£1.11	£1.66

The sample statistics suggest a difference in the average amount of pocket money in the two schools (with School 2 having the higher sample mean), but there also appears to be considerable variation in our samples – both have standard deviations greater than £1. If we took new samples of 60 children from each school, our new sample means would almost certainly be different. We want to know, based on the sample data we have collected, if we have strong evidence of a difference in average pocket money in the two schools. Interpreting the p-value from a null-hypothesis test of significance will help us to do this.

Table 2: Statistical output from R for t-test of means

```
Welch Two Sample t-test

data:  school2 and school1
t = 2.0071, df = 102.796, p-value = 0.047
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
 0.006117296 1.027216038
sample estimates:
mean of x mean of y
 5.400000  4.883333
```

We conduct a t-test using the free statistical software 'R'. The results are shown in Table 2 (left) with the p-value highlighted. We can represent the meaning of this p-value (*p*) as follows:

$$p = \Pr\left(Data \middle| H_0\right)$$

In words, this means: **the p-value is equal to the probability of sampling the data we sampled [or more extreme data] given that the null hypothesis is true.**

So, what is our null hypothesis? We define the null hypothesis when choosing and conducting a test. In this example, the null hypothesis is "there is **no difference** in average pocket money in the two schools" (i.e. the difference in mean pocket money is equal to zero). The p-value output from our test here is 0.047 or 4.7%. So, the test tells us that the probability of sampling the data we sampled [or more extreme data] given that there is no difference in average pocket money in the two schools is 4.7%. Our sample difference is clearly quite unlikely given that the null-hypothesis is true, therefore we decide to reject our null hypothesis in favour of an alternative: that there is an actual difference in average pocket money in the schools.

However, this does not mean that there *definitely* is a difference in mean pocket money between the two schools. Statistical tests rely on probabilities not absolutes, so assertions based on their results are not strictly valid. However, they enable us to quantify the probability of sampling data with this magnitude of difference if a difference did not actually exist in the population. This is the key to interpreting p-values in null hypothesis tests of significance.

Often in social research, we use the 5% p-value as a boundary at which to assume we have evidence to reject the null hypothesis, i.e. if $p \leq 5\%$, the difference in our samples is unlikely enough, given the null hypothesis, that we infer there is a difference in the population; if $p > 5\%$ we do not think there is sufficient evidence of a difference in the population and stick with the null hypothesis. This boundary is, of course, arbitrary and it makes sense to interpret data with intelligence and reflection rather than making decisions with a simplistic rule. Indeed, in some applications, $p \leq 10\%$ is deemed acceptable, in others 0.1% is used. Also, sample size affects the p-value, with larger sample sizes tending to give smaller p-values, so it makes sense to interpret them along with the difference (or effect size) measured in your sample data.