Controversies regarding significance (hypothesis) testing

We know that hypothesis testing is fundamental to inferential statistics because it allows us to make probabilistic statements about data. It defines a procedure that controls the probability of incorrectly *deciding* that a null hypothesis Ho (a default position) is incorrect.   This is based on how likely it would be for a set of observations to occur if the null hypothesis were true.

However, because those statements are probabilistic rather than absolute, the possibility of error is inherent in the whole process. Statisticians have defined two types of error possible when making decisions using inferential statistics and have established levels for error rates that are commonly considered acceptable. These errors are called Type I (false positive) and Type II (false negative) errors.

For example: in the evaluation of detection limit in chemical trace analysis, when we state "Not Detected (ND) " for a measurand in our test report, it indicates that this measurand may be present below the detection limit and is not completely absent in the sample analysed. But, we do not wish to claim the presence of the measurand when it is actually absent; i.e. a false positive or Type I (α) decision error.   Equally, we do not wish to report that the measurand is absent when it is truly present, i.e. a false negative or Type II (β) decision error. Hence, in any situation, we might make a correct decision or we might commit a Type I or Type II error.

The level of acceptability for Type I error is conveniently set at 0.05 in modern statistics, which means that we accept a 5% probability of Type I error. To put it another way, we understand when setting the alpha level at 0.05 in our study, we have a 5% chance of rejecting the null hypothesis when we should fail or reject it.

However, the use of hypothesis testing has not gone unchallenged in the past many years, particularly on the universal choice of $\alpha = 0.05$ significance level, amongst many other criticisms. One of the many critics is Jacob Cohen, whose arguments are presented in his 1994 article titled "The Earth is Round ($p < 0.05$)", which can be found at:
http://www.ics.uci.edu/~sternh/courses/210/cohen94_pval.pdf

There are criticisms on the wisdom of taking the 0.05 value. We agree that we must establish some standard for statistical significance to minimize the possibility of attributing significance to differences due to sampling error or other chance factors, but why this 0.05 magical level?

The significance level of results calculated on a sample is affected by many factors, including the size of the sample involved, and overemphasis on the p-value of a result ignores the many reasons a particular study may or may not have found significance. It is a common saying among statisticians that if you have a large enough sample, even a tiny effect will be statistically significant.

However, a large sample size can inevitably reduce the unreliable and invalid part of the variance in the measurements and narrow the confidence limits of the reported result. It is generally believed that confidence intervals contain all the information to be found in significance tests and much more. Many researchers do not like to state confidence limits or intervals in their conclusions because they can be embarrassingly large.

The take-home message therefore is that statistical methods are powerful tools but they do not relieve researchers of the need to use their common sense as well. In other words, the choice of a significance level in a study is entirely yours.