A common complaint about Hypothesis Testing: The choice of the significance level $\alpha$ is essentially arbitrary.

To make matters worse, the conclusions (Reject or Don’t Reject $H_0$) depend on what value of $\alpha$ is used.

Eg: In the "Quarter Pounders" example, we had $H_0: \mu = .25$, $H_1: \mu < .25$, $\bar{x} = .24$ and $z = -2.02$.

Since $-z_{.05} = -1.645$, we have enough evidence to reject $H_0$ at level .05.
Since $-z_{.01}=-2.326$, we do not have strong enough evidence to reject $H_0$ at level .01.

A Key Question: Exactly how strong is the statistical evidence (provided by our given data) against $H_0$?

A reasonable answer is given by the $p$-value: The smallest level at which $H_0$ would be rejected.

**Eg:** In the "Quarter Pounder" example, we were able to reject $H_0$ at level .05. But our actual evidence, as contained in our observed $z$ statistic, $z_{obs}=-2.02$, is somewhat stronger. In fact, if we did a test with $\alpha$ somewhat smaller than .05, we
would still be able to reject $H_0$. By just reporting that we can reject at level .05, we are throwing out important information.

Instead, let’s find the smallest level $\alpha$ at which we could do a test and still reject $H_0$. This smallest $\alpha$ is the $p$-value. It measures the strength of the strongest case against $H_0$ which can be supported by the data.

To compute $p$ for this example, imagine what happens in a (hypothetical) level $\alpha$ test as $\alpha$ is decreased from .05. At first, you will still be able to reject $H_0$ because the critical value $-z_\alpha$ will be to
the right of the observed $z$-value, $z_{obs} = -2.02$. But as $\alpha$ is decreased, the critical value moves to the left. After $\alpha$ gets so small that $-z_\alpha$ has crossed to the left of $-2.02$, we will no longer be able to reject $H_0$. So the smallest $\alpha$ for which $H_0$ can be rejected (i.e., the $p$-value) is the $\alpha$ for which $-z_\alpha = -2.02$.

Thus, the $p$-value is the level of a test for which the critical value is $-2.02$. Since the level of a test is the probability of rejecting $H_0$ when $H_0$ is true, our $p$-value is the area to the left of $-2.02$ under the standard normal curve. So we get $p = .0217$. This is between .05 and .01, as expected.
Alternative Definition of \( p \)-Value: The \( p \)-value is the probability that you would obtain evidence against \( H_0 \) which is at least as strong as that actually observed, if \( H_0 \) were true.

If properly interpreted, this is equivalent to the earlier definition. Note that the precise meaning of "evidence" and "at least as strong" depends on which test statistic is being used, and on the exact nature of the alternative hypothesis (e.g., one-sided or two-sided).

The following should be clear from the alternative definition given above:
The smaller the $p$-value, the stronger the evidence that $H_0$ is false.

More informally, we can think of the $p$-value as an index of "flukiness" (unusualness) of the sample, assuming $H_0$ is true, or as the level of "reasonable doubt" of $H_1$. Note, however, that the $p$-value is NOT the probability that $H_0$ is true.

**Eg:** Your company’s brand awareness, as measured on a scale of 1 to 10, was found to have increased an average of .3 points when 200 people were shown an advertisement and questioned before and after. The standard deviation of the increase was
1.39 points. Is this a significant change?

**Sol:** If \( \mu \) is the mean increase for all consumers in the target population, the null hypothesis of no increase is \( H_0 : \mu = 0 \).

The alternative hypothesis that the ad does have a positive effect is \( H_1 : \mu > 0 \).

(Note, however, that if a significant *decrease* in awareness is also important, you should be using a two-sided alternative. One-sided alternatives should only be used if, before you even look at the data, you are sure that you and your bosses are only interested in or worried about one type of deviation from \( H_0 \).)
Let’s compute the $p$-value. We have $n = 200$, $\mu_0 = 0$, $\bar{x} = .3$, $s = 1.39$ and $t = 3.05$. Since $n$ is "large", $t$ would have a standard normal distribution if $H_0$ were true. Thus, the $p$-value is the probability that a standard normal would exceed 3.05, that is, $p = .0011$.

Most analysts would agree that the increase is highly significant. If the ad had no effect, the probability of getting such good results as obtained here would be only 11 out of 10,000, so it would be quite a statistical "fluke" to get such an unusual sample. Instead, it is much more plausible that the ad actually was effective.
Comment #1: If we had used a two-sided alternative, the $p$-value would be .0022, twice what it was before, since "strong evidence" would mean that $t$ is either extremely positive or extremely negative.

Comment #2: Even though the increase was highly significant in a statistical sense, it wasn’t very much in a practical sense. (Only .3 on a scale of 10.) When you report a statistically significant change, you are really just saying that the change was statistically detectable, i.e., more than could be plausibly be accounted for by natural variability alone.
Relationship Between $p$-Values and Hypothesis Tests

Computing a $p$-value is not the same thing as performing a hypothesis test. The key differences are:

(1) In hypothesis testing, we use a fixed (arbitrary) value of $\alpha$, such as .05 or .01. The only possible conclusions are: Reject $H_0$ or Do Not Reject $H_0$, at level $\alpha$.

In reporting a $p$-value, we are not performing a specific test.

Instead, the $p$-value tells us what the conclusions would be for all possible tests which could be performed: $H_0$ would be rejected for any test whose level exceeds the $p$-value.
So the $p$-value is much more informative than the results of any particular test.

(2) $p$-values provide the relevant information for the decision maker, but do not dictate any particular decision. Instead, the final decision is left to the analyst, who may wish to consider other factors besides the numerical data as summarized by the $p$-value.

So the $p$-value allows for a much more flexible framework for making informed decisions than can be obtained by rigid hypothesis testing.