

Why a P -value includes the probability of outcomes farther from the null hypothesis than the observed outcome.

Some of you have wondered why a P -value is the probability (given that H_0 is true) not only of the observed value of the test statistic but of all possible more-extreme values of the statistic. This seems particularly unclear when considering a test with a discrete test statistic, such as the binomial test used in the recent homework problem to determine whether the old farmer and his forked stick could detect the presence of water. Here's an attempt at an explanation.

Data (= values of a test statistic) that are far [in the direction(s) specified by the H_a] from what is expected under H_0 provide strong evidence against H_0 — this much is common sense. In the homework question, if the farmer had been correct for all 5 barrels this would have been even stronger evidence against H_0 than was the observation of 4 out of 5 correct. For this reason a test with a fixed significance level consists not of a “critical point” but of a “critical region”: a range of values, starting at a “critical point” but also including all points farther from H_0 , such that whenever the test statistic falls in this region the null hypothesis is rejected. The size of this critical region is determined in order to give the desired risk of Type I error, i.e. α ; this is the probability, given that H_0 is true, of observing a value of the test statistic anywhere in the critical region. In other words, the “significance” measured by α refers to the process used in the test, which is defined by the critical region.

A P -value can be regarded as the greatest significance level (smallest α) at which the null hypothesis would be rejected given the observed data; this is simply a re-statement of the definition that we reject H_0 at significance level α when $P \leq \alpha$. This then implies that a P -value is the Type I error rate of a test which would use the value of the test statistic obtained in the particular study as its critical point for rejecting H_0 . Since α is the probability of a test statistic being in the critical region given H_0 , so then must the P -value include the probability of the more extreme possible results as well as of the observed result.

To be more concrete, suppose we did a lot of tests in the binomial situation, with $n = 5$ and $H_0: p = 0.5$, and for every test used $X \geq 4$ as the critical region. What then is the significance level, α , or Type I error rate, of these tests? It is the probability, given that the null hypothesis is true, of getting $X = 4$ or $X = 5$, which is $P = 0.1876$. If we wanted a smaller α , we would need a smaller critical region, not including $X = 4$, and so would not reject on the basis of observing 4 correct answers. Conversely, if we used a larger α , we would have a larger critical region and therefore would still reject for $X = 4$. So the P -value, the smallest α at which we would reject the null hypothesis, is 0.1876: the probability (assuming the null hypothesis is true) of the test statistic being as extreme or more extreme than was observed.