e-values in Hypothesis Testing

Bao Han Ngo

Winter 2025 DRP

p-values

Hypothesis testing with p-values is commonly used in research to reject a null hypothesis. A p-value is a random variable P such that $P_{H_0}(P \leq \alpha) \leq \alpha$ under the null hypothesis. A small p-value would indicate that there is a low probability of seeing such data under the null, and thus a small p-value represents evidence against the null. Using significance level α, if a p-value is below α, we would reject the null hypothesis. One could understand the "p" in p-value to stand for probability.

There are some caveats to constructing statistically valid p-values. When working with p-values, is important to specify both α and the sample size before conducting any testing. Additionally, combining dependent p-values from different experiments can be difficult. The statistician should also know the data collection procedure in order to calculate the p-value properly.

e-values

e-values are an alternative to p-values in hypothesis testing and offer many advantages. An e-value is a random variable E with an expected value of at most 1 under the null hypothesis, $E_{H_0}[E] \leq 1$. Whereas the "p" in p-value stands for probability, the "e" in e-value stands for expectation. The "e" can also stand for evidence as a larger e-value represents more evidence against the null. By Markov's inequality, $\frac{1}{e}$ is a valid p-value, meaning that the null can be rejected if the e-value exceeds $\frac{1}{\alpha}$. e-values are more conservative than p-values so observing a p-value of 0.01 does not necessarily mean that the e-value will be $100 = \frac{1}{0.01}$.

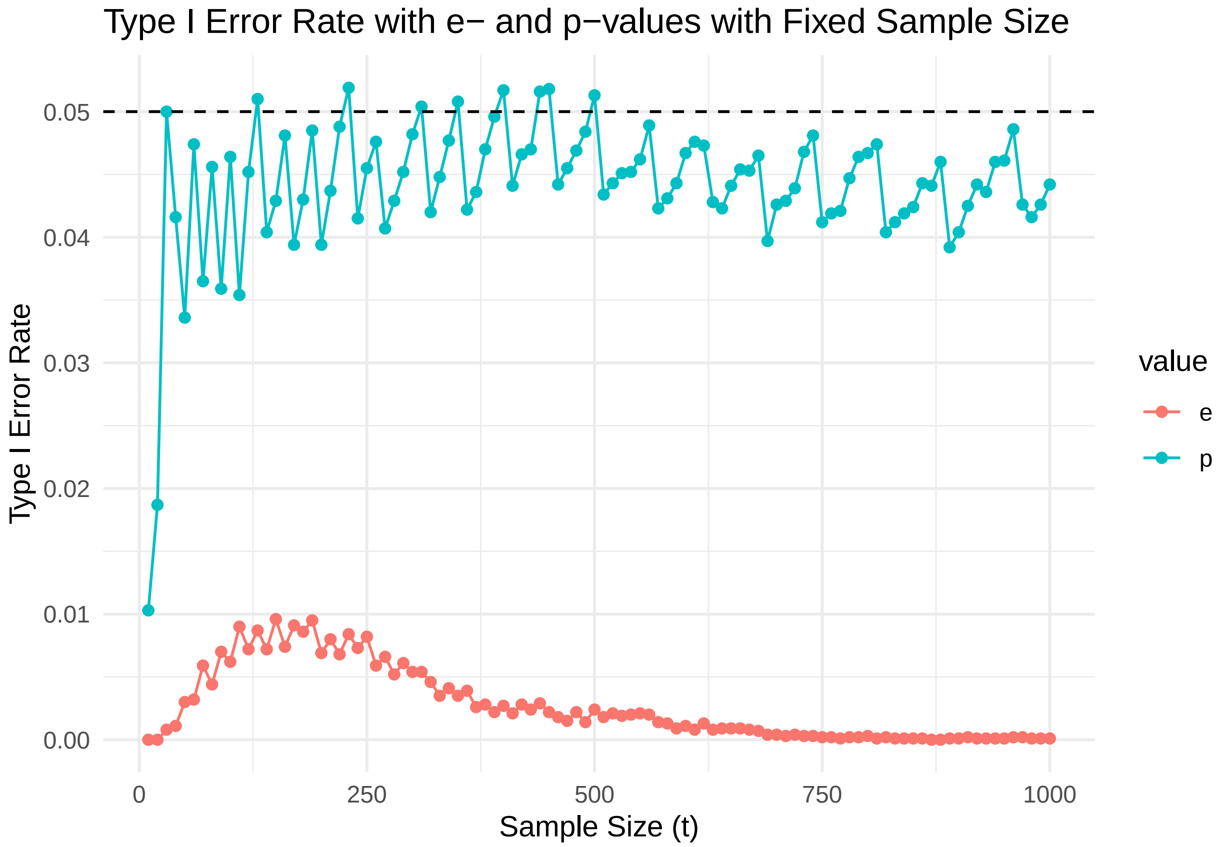
Compared to p-values, α can be chosen after the test when using e-values. Additionally, e-values do not require a pre-specified sample size before testing. This allows you to conduct tests as the data comes in rather than having to wait until the sample is fully collected. If an early test shows significance, it is valid to reject the null immediately with no need to finish collecting the rest of the data. With e-values being expectations, it is more straightforward to combine e-values from different experiments, even when they might be dependent. Moreover, e-values can still be calculated even if the data collection procedure is unknown. Finally, there are just some problems in testing that p-values cannot be used for, but e-values can be (for example, testing if data comes from two Gaussians).

An example of an e-value can be constructed from a likelihood ratio. For hypotheses, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, the likelihood ratio would be $\Lambda = \frac{L(\theta_0|x)}{L(\theta_1|x)}$. Taking the inverse, $\frac{1}{\Lambda} = \frac{L(\theta_1|x)}{L(\theta_0|x)}$ has an expected value of 1 and is thus an e-value.

Hypothesis Testing with p- and e-values for Fixed Sample Sizes

In a valid $\alpha-$level test, we expect to commit a type I error with probability α. In this simulation, we generate t samples from a $Bernoulli(p = 0.5)$. For every sample size t, we calculate a p-value and e-value to test (and maybe reject) $H_0 : p = 0.5$ at $\alpha = 0.05$ significance. If we reject the null, that would be a type I error. In 10,000 iterations, type I errors should be committed at most 5% of the time if the test is statistically valid.

The p-value case has an alternative hypothesis of $H_1 : p > 0.5$ while the e-value case has an alternative hypothesis of $H_1 : p = 0.6$. It is possible to have a composite alternative hypothesis when using e-values (requires splitting the data into two groups so that one group may be used to pick an alternative hypothesis), but for simplicity, a simple alternative was chosen for the e-value case. Either way, the same null hypothesis is being tested. The p-value is calculated using the binomial tail probability and the e-value is calculated using the inverse of the likelihood ratio, $e = \frac{0.6^x(1-0.6)^{t-x}}{0.5^x 0.5^{t-x}}$, where x is the sum of successes in t trials.



As expected, with a pre-specified sample size, the type I error rate is on average 0.05 or below with both p- and e-values.

Hypothesis Testing with p- and e-values with Early Stopping

Next, we add early stopping into the simulation by calculating a p- and e-value after every 10 samples, up to 1000 samples. If the test leads to a rejection, we stop testing collecting data and stop testing. Otherwise, we conduct another test with the addition of the next 10 points. With so many tests being run, there could be more opportunities to commit a type I error. At the of of the simulation, we calculate the proportion of iterations that had a rejection of the null using either p- or e-values.

The results show that with p-values, rejection occured in 26.2% of the iterations. On the other hand, rejection occurred in only 3.7% of the iterations when using e-values. With p-values, the type I error rate has far exceeded α when we "peeked" at our data and conducted tests before our full sample size was collected. This shows that p-values do not maintain validity when multiple tests are conducted before obtaining the full pre-specified sample size. However, e-values do maintain validity, allowing for peeking at the data, conducting tests without pre-specifying the sample size, and early stopping.

An Example

Suppose you are a statistician testing $H_0 : p = 0.5$ (against $H_0 : p > 0.5$ in the p-value case, $H_1 : p = 0.6$ in the e-value case) at $\alpha = 0.05$. Since we are designing the simulation, we know that the data comes from $Bernoulli(p = 0.6)$, but the as a statistician in the real world, you wouldn't know that. Therefore, being the good statistician you are, you pre-specify a sample size of $n = 200$ so that you are able to conduct a test with a p-value. After putting in lots of time to collect all 200 datapoints, you calculate $p_{n=200} = 0.01$ and reject the null (which you should, since the data does not come from the null distribution). If you were a bad statistician and conducted a test partway through data collection before reaching your pre-specified 200 observations, you would have realized that your first significant p-value would have appeared after only 93 points ($p_{n=93} = 0.048$). But, you practice valid statistics and waited until all 200 observations were collected to calculate one single p-value at the end.

If you had elected to instead conduct testing with e-values, it would have been valid to calculate an e-value on just part of the data, even before colleting all 200 points. In this case, you would have realized that after 157 observation, you get an e-value of $e_{n=157} = 21.8 > \frac{1}{\alpha}$, allowing you to stop data collection early and reject the null. You could have saved a lot of time and money by not collecting those additional 43 points. Only in the e-value case, is this valid to do! The fact that the first significant e-value comes later than the first significant p-valie ($157 > 93$) illustrates the conservative nature of e-values compared to p-values.

Conclusions

p-values are being abused across disciplines and the abuse of p-values can lead to highly inflated error rates. e-values are one solution to address this abuse while also offering advantages when it comes to study design, data collection, and testing.