Lecture 1: p-values and e-values

Eugenio Clerico

November 2025

Testing with p-values. In its simplest formulation p-value testing can be summarised as follows. First, one fixes a significance level $\alpha \in [0,1]$. Then picks a p-value p and evaluates it on the data X. Finally, the null hypothesis $\mathcal{H}$ is rejected (at level α) if $p \leq \alpha$, otherwise the hypothesis is not rejected, as there is not enough evidence to do so. Why does this work? Essentially by the definitions itself of significance level and p-value.

Let $\mathcal{X}$ be the data space (namely $x \in \mathcal{X}$ is a data set that can be used for a test). Fix a hypothesis $\mathcal{H}$, that is a set whose elements are probability distributions on $\mathcal{X}$. A test has *significance level* α if it can reject a dataset coming from any $Q \in \mathcal{H}$ with probability at most α. Formally,

$$\sup_{Q \in \mathcal{H}} Q(\text{the test rejects } \mathcal{H}) \leq \alpha\,.$$

A *p-variable* is a measurable function $p : \mathcal{X} \to [0,\infty]$, such that for all $\alpha \geq 0$ we have

$$Q(p \leq \alpha) \leq \alpha\,,$$

for every measure $Q \in \mathcal{H}$.[1] The term *p-value* refers to the actual value that a p-variable takes on the observed data. When testing with a p-value, after having fixed a level α, the test rejects $\mathcal{H}$ precisely when $p(X) \leq \alpha$. The significance level guarantee is ensured by the fact that

$$\sup_{Q \in \mathcal{H}} Q(\text{reject } \mathcal{H}, \text{ testing with } p \text{ with threshold } \alpha) = \sup_{Q \in \mathcal{H}} Q(p \leq \alpha) \leq \alpha\,.$$

α-hacking. By design, α-testing with p-values requires to first fix a level α and a p-variable p (namely, *first fix a test*), then look at the data and evaluate the p-value. Yet, there are cases where this can feel like a limitation. Say that $\alpha = .05$ is fixed before the experiment. Then p is measured and a p-value of .001 is observed. Declaring a significance level of .05, when such a small p-value has been observed, feels a bit like a waste! One could have declared a much smaller α (say $\alpha = .01$) and still the test would have found enough evidence to reject the null, and the conclusion would have looked much stronger. Yet, deciding the significance level *a posteriori* is a practice (α-hacking) not allowed by the standard p-value testing.

To understand what goes wrong with α-hacking, let us consider the following example. We are given a hypothesis $\mathcal{H}$ and a dataset X. Let us assume that we know an exact p-variable p, namely under the null we have that $p(X)$ is uniformly distributed. We follow the following procedure.

1. We evaluate $p(X)$.
2. If $p(X) \leq .01$, then we reject $\mathcal{H}$, declaring significance level $\alpha = .01$.
3. If $p(X) > .01$, but still $p(X) \leq .05$, then we reject $\mathcal{H}$, but declaring significance level $\alpha = .05$.
4. If $p(X) > .05$, we give up, and we say that there is not enough evidence to reject the null (declaring $\alpha = .05$ as the used threshold).

Now, recall that the meaning of the significance level α is that "*The probability of rejecting under a true null is at most* α.". But this is false if we are in the first case, where we have declared $\alpha = .01$. Indeed, our procedure rejects with 5% chances if the data are actually coming from a $Q \in \mathcal{H}$, as the rejection region is $\{p(X) \leq .05\}$, and $p(X)$ is uniformly distributed on $[0,1]$. Surely, if we had observed a

[1] Note that often p-variables are defined to be valued in $[0,1]$, and the condition is required to hold for every $\alpha \in (0,1)$. Yet, allowing values larger than 0 does not have an impact on testing and is convenient for what will follow here.

p-value of .02, we would have declared α equal to .05 and not .01, but the definition of significance level does not care about what level one is declaring, just about the binary outcome *Reject/Don't reject*.[2]

What is going wrong? In short, p-value testing simply does *not* allow data-dependent choices of the significance level (in other words: *α must be fixed before looking at the data*). The definition of a p-variable requires that

$$Q(p \leq \alpha) \leq \alpha$$

holds for every *fixed* threshold α. Nothing in the definition guarantees the same inequality when α can possibly depend on the data.

Conditional validity. If one wishes to preserve a meaningful statistical guarantee while allowing post-hoc tuning of α, then the requirement on the p-variable must be strengthened accordingly. More explicitly, suppose we allow the significance level to be chosen after observing the data, through some random variable $\hat{\alpha} : \mathcal{X} \to [0, \infty]$. In order for p-value testing to remain valid under such a rule, one would have to require that, for every $Q \in \mathcal{H}$,

$$Q\big(p \leq \hat{\alpha} \,\big|\, \hat{\alpha}\big) \leq \hat{\alpha}\,, \quad Q\text{-almost surely.} \tag{1}$$

In words: *conditionally* on the declared significance level, the probability (under the null) that the p-variable falls below that level must still be at most the level itself. If a p-variable satisfies this stronger requirement, then one could indeed give a clear statistical meaning to a significance level $\hat{\alpha}$ chosen in a data-dependent manner, by saying that a test has significance level $\hat{\alpha}$ if, for all $Q \in \mathcal{H}$,

$$Q\big(\text{the test rejects } \mathcal{H} \,\big|\, \hat{\alpha}\big) \leq \hat{\alpha}\,, \quad Q\text{-almost surely.}[3]$$

So, are we done? Can we just say that if we take any random variable $p : \mathcal{X} \to [0, \infty]$ satisfying (1) and we use it for our test we are allowed to let the significance level depend on the data? Sure thing! So, are we done? Well...

The problem now is that for the testing purpose there is essentially no *useful* p that satisfies (1), as it implies that $p \geq 1$, Q-almost surely, for every $Q \in \mathcal{H}$. To see this, fix any $\varepsilon \in (0, 1)$ and let $\hat{\alpha} = \varepsilon$, if $p \leq \varepsilon$, and 1 otherwise. Then, if there is $Q \in \mathcal{H}$ such that $Q(p \leq \varepsilon) > 0$, on $\{p \leq \varepsilon\}$ (namely, where $\hat{\alpha} = \varepsilon$) we have

$$Q(p \leq \hat{\alpha} \,|\, \hat{\alpha} = \varepsilon) = Q(p \leq \varepsilon \,|\, p \leq \varepsilon) = 1 > \varepsilon\,.$$

It follows that $Q(p \leq \varepsilon) = 0$ for every $\varepsilon \in (0, 1)$, and so $Q(p < 1) = 0$.

So, we should not expect post-hoc validity to hold in the strong sense implied by (1) for p-variables. Still, one may wonder whether a weaker form of validity could be recovered. That is, although the exact significance-level guarantee cannot be maintained once α becomes data-dependent, perhaps a controlled relaxation of the requirement is still possible.

Expected conditional validity. The idea is to quantify how much the validity condition is *violated* when $\hat{\alpha}$ depends on the data. Concretely, we measure by how much $Q(p \leq \hat{\alpha} \mid \hat{\alpha})$ may exceed $\hat{\alpha}$. A natural way to capture this discrepancy is through the ratio

$$\frac{Q(p \leq \hat{\alpha} \mid \hat{\alpha})}{\hat{\alpha}}\,,$$

with the convention that $0/0 = 1$ whenever needed. From the earlier discussion we have seen that we cannot hope to control this ratio almost surely under the null, uniformly over all choices of $\hat{\alpha}$. However, we might try to relax the requirement and only control it *in expectation*. Namely, let us attempt to define a *post-hoc p-variable* as a random variable $p : \mathcal{X} \to [0, \infty]$ such that

$$\sup_{Q \in \mathcal{H}} \sup_{\hat{\alpha}} \mathbb{E}_Q\left[\frac{Q(p \leq \hat{\alpha} \mid \hat{\alpha})}{\hat{\alpha}}\right] \;\leq\; 1\,, \tag{2}$$

[2] This is a subtle but important point. The definition of significance level that we are considering is meaningful for binary tests! If we consider a test with three possible outcomes: *Don't reject*, *Reject declaring* $\alpha = .01$, and *Reject declaring* $\alpha = .05$, and we say that the significance level does now represent the an upper bound on the probability to declare the specific outcome under the null, then it would be true that any p-variable ensures that the outcome *Reject declaring* $\alpha = .01$ is compatible with a significance level 0.01 and the outcome *Reject declaring* $\alpha = .05$ is compatible with a significance level 0.05. Yet, this requires a different and non-standard interpretation of the meaning of significance level.

[3] Note that, to define a data-dependent significance level, it would not really make sense to require that, for all $Q \in \mathcal{H}$, $Q(\text{rejection}) \leq \hat{\alpha}$, since the left-hand side is a deterministic quantity while the right-hand side is a random variable.

where the inner supremum runs over all random variables $\hat{\alpha} : \mathcal{X} \to [0, \infty]$.

Before going on and checking whether this actually leads to a practically usable notion, let us pause for a moment. Does it make sense to look at things in expectation? After all, the usual motivation for p-values is that they provide a high-probability guarantee on the statistical fluctuation of the test outcome, which might not really combine well with guarantees in expectation. However, one needs to remark that the standard definition of p-values can also be written in an expectation form. Indeed, for any fixed $\alpha \geq 0$ and any $Q \in \mathcal{H}$,

$$Q(p \leq \alpha) \leq \alpha \qquad \Longleftrightarrow \qquad \mathbb{E}_Q\left[\frac{\mathbf{1}_{p \leq \alpha}}{\alpha}\right] \leq 1\,.$$

In the post-hoc setting we are essentially doing the same, but with a data-dependent level. Indeed, we have

$$\mathbb{E}_Q\left[\frac{Q(p \leq \hat{\alpha} \mid \hat{\alpha})}{\hat{\alpha}}\right] = \mathbb{E}_Q\left[\frac{\mathbf{1}_{p \leq \hat{\alpha}}}{\hat{\alpha}}\right], \tag{3}$$

and we require this to be at most 1, uniformly over all choices of $\hat{\alpha}$. With this in mind, the proposed condition can be viewed as a natural way to strengthen the standard definition of p-variable.

Post-hoc p-variables. We defined a *post-hoc p-variable* as a random variable $p : \mathcal{X} \to [0, \infty]$ that satisfies (2). At first sight, this looks like a rather complicated condition. Can we find a simpler characterisation?

Using (3), we see that we need to look at

$$\sup_{\hat{\alpha}} \mathbb{E}_Q\left[\frac{\mathbf{1}_{p \leq \hat{\alpha}}}{\hat{\alpha}}\right],$$

where the supremum is taken over all random variables $\hat{\alpha} : \mathcal{X} \to [0, \infty]$.

The key observation is the following (easily verified) pointwise identity: for any fixed $x \in \mathcal{X}$,

$$\sup_{\alpha \geq 0} \frac{\mathbf{1}_{p(x) \leq \alpha}}{\alpha} = \frac{1}{p(x)}\,,$$

with the convention that $0/0 = 1$ and $1/0 = \infty$. Indeed, if $\alpha < p(x)$ the ratio is 0, while if $\alpha \geq p(x)$ it equals $1/\alpha$, which is maximised by choosing $\alpha = p(x)$.

This immediately yields, for any fixed $Q \in \mathcal{H}$,

$$\sup_{\hat{\alpha}} \mathbb{E}_Q\left[\frac{\mathbf{1}_{p \leq \hat{\alpha}}}{\hat{\alpha}}\right] \leq \mathbb{E}_Q\left[\sup_{\alpha \geq 0} \frac{\mathbf{1}_{p \leq \alpha}}{\alpha}\right] = \mathbb{E}_Q\left[\frac{1}{p}\right],$$

and equality is attained by the particular choice $\hat{\alpha} = p$, for which

$$\mathbb{E}_Q\left[\frac{\mathbf{1}_{p \leq \hat{\alpha}}}{\hat{\alpha}}\right] = \mathbb{E}_Q\left[\frac{\mathbf{1}_{p \leq p}}{p}\right] = \mathbb{E}_Q\left[\frac{1}{p}\right].$$

Therefore,

$$\sup_{\hat{\alpha}} \mathbb{E}_Q\left[\frac{Q(p \leq \hat{\alpha} \mid \hat{\alpha})}{\hat{\alpha}}\right] = \sup_{\hat{\alpha}} \mathbb{E}_Q\left[\frac{\mathbf{1}_{p \leq \hat{\alpha}}}{\hat{\alpha}}\right] = \mathbb{E}_Q\left[\frac{1}{p}\right].$$

In particular, the post-hoc validity condition (2) is equivalent to the much simpler requirement

$$\sup_{Q \in \mathcal{H}} \mathbb{E}_Q\left[\frac{1}{p}\right] \;\leq\; 1\,.$$

E-variables. Let us quickly summarise some of the definitions that we have discussed so far. A p-variable for $\mathcal{H}$ is a non-negative random variable p that satisfies

$$\sup_{Q \in \mathcal{H}} \sup_{\alpha \geq 0} \mathbb{E}_Q\left[\frac{\mathbf{1}_{p \leq \alpha}}{\alpha}\right] \leq 1\,.$$

A post-hoc p-variable is a non-negative random variable p that satisfies

$$\sup_{Q \in \mathcal{H}} \mathbb{E}_Q\left[\sup_{\alpha \geq 0} \frac{\mathbf{1}_{p \leq \alpha}}{\alpha}\right].$$

Equivalently, a post-hoc p-variable is a non-negative random variable p that satisfies

$$\sup_{Q \in \mathcal{H}} \mathbb{E}_Q \left[\frac{1}{p} \right] \leq 1 \,.$$

A non-negative random variable whose expectation under the null is upper bounded by 1 is exactly the definition of an *e-variable*. So, we can define a post-hoc p-variable as: *A post-hoc p-variable is a non-negative random variable p whose inverse $1/p$ is an e-variable.*

Also note that the inverse of an e-variable E is always a p-variable! Indeed, by Markov's inequality we have that

$$Q\left(\frac{1}{E} \leq \alpha \right) = Q\left(E \geq \frac{1}{\alpha} \right) \leq \frac{\mathbb{E}_Q[E]}{1/\alpha} \leq \alpha \,.$$

In particular, since if $1/E = p$ then $1/p = E$ is an e-variable, we see that the inverse of an e-variable is always a post-hoc p-variable. So, we can also define an e-variable as: *An e-variable is a non-negative random variable E whose inverse $1/E$ is a post-hoc p-variable.*

This shows that there is a one-to-one canonical correspondence between e-variables and post-hoc p-variables: the bijection $E \mapsto 1/E$. In a way, this shows that one can identify the e-variables with a subset of the p-variables, under the canonical injection $E \mapsto 1/E$.

Example: back to α-hacking. Let us go back to the initial example that we discussed when introducing the α-hacking. We have a uniformly distributed p-variable p under Q. This is not a post-hoc p-variable, as $\mathbb{E}_Q[1/p] = \infty$. So, if we set the significance level depending on the data we are not assured that the violation of the conditional validity is bounded by 1 in expectation. For instance, if we set

$$\hat{\alpha} = \begin{cases} .01 & \text{if } p \leq .01; \\ .05 & \text{otherwise,} \end{cases} \tag{4}$$

we have that

$$\mathbb{E}_Q \left[\frac{Q(p \leq \hat{\alpha} | \hat{\alpha})}{\hat{\alpha}} \right] = .01 \times \frac{1}{.01} + .99 \times \frac{.04/.99}{.05} = 1.8 > 1 \,.$$

Define

$$p' = \begin{cases} .01 & \text{if } p \leq .002; \\ .05 & \text{if } p \in (.002, .042]; \\ \infty & \text{otherwise.} \end{cases}$$

This is a post-hoc p-variable, since we have

$$\mathbb{E}_Q \left[\frac{1}{p} \right] = .002 \times 100 + .04 \times 20 = 1 \,.$$

We can indeed use $\hat{\alpha}$ in (4) as significance level when testing with p', and we have the guarantee

$$\mathbb{E}_Q \left[\frac{Q(p' \leq \hat{\alpha} | \hat{\alpha})}{\hat{\alpha}} \right] = .01 \times \frac{Q(p' \leq .01 | p \leq .01)}{.01} + .99 \times \frac{Q(p' \leq .05 | p > .01)}{.05} = \frac{1}{5} + \frac{3.2}{5} = .84 < 1 \,.$$

In practice, under Q we have that this procedure rejects overall if the original p is at most .042 (so it is more conservative than just testing with p with the α-hacking procedure, as expected). When $p \leq .002$ it declares significance level .01, while if $p > .02$ it declares .05.

As a final remark, one could also use a data-dependent significance level $\hat{\alpha}'$ defined as .01 if $p' \leq .01$, and .05 otherwise. One can easily check that this would bring an expected violation of the conditional significance level bounded by 1, but it is a less strong procedure, as we always have $\hat{\alpha}' \geq \hat{\alpha}$.

Calibrators. We have seen that passing from an e-variable to a p-variable is rather easy: one can simply apply the mapping $E \mapsto 1/E$. In the opposite direction, however, the situation is more delicate. If p is a post-hoc p-variable, then the transformation $p \mapsto 1/p$ indeed yields an e-variable. But this will *not* be the case when p is only a standard p-variable and does not satisfy post-hoc validity. In general, there is no simple universal way to turn an arbitrary p-variable into an e-variable.

Rather than trying to construct completely general mappings from p-variables to e-variables (or, equivalently, to post-hoc p-variables) for a given hypothesis, the literature usually takes a simpler and

more manageable route, centred around the notion of a *calibrator*. A calibrator is a non-increasing function $f : [0, \infty] \to [0, \infty]$ such that, for *any* $\mathcal{X}$, *any* hypothesis $\mathcal{H}$ on $\mathcal{X}$, if p is a p-variable then

$$E = f \circ p$$

is an e-variable. The monotonicity requirement simply matches the fact that stronger evidence against the null corresponds to larger e-values and smaller p-values. Focusing on such structured transformations keeps things clean and workable. This is why calibrators have become the standard way of turning e-values into p-values in a principled and interpretable manner.

It is possible to give an explicit characterisation of all the calibrators: a measurable $f : [0, \infty] \to [0, \infty]$ is a calibrator if, and only if,

$$\int_0^1 f(u)\mathrm{d}u \leq 1\,. \tag{5}$$

In particular, some examples of calibrators are $f(u) = \max\{0, 2(1-u)\}$, $f(u) = \log \frac{1}{\min\{u,1\}}$, and $f(u) = k \max\{u^{k-1}, 1\}$ for $k \in (0, 1)$. Note that the first one is very conservative, as it can never reject if $\alpha < 1/2$!

Let us prove that (5) holds if f is a calibrator. We can pick $\mathcal{X} = [0, 1]$ and $\mathcal{H} = \{\mathrm{Unif}\}$, the uniform distribution. Then $p : x \mapsto x$ is a p-variable and so $E = f \circ p$ is an e-variable, since f is a calibrator. This means precisely that

$$1 \geq \int_{\mathcal{X}} E(x)\mathrm{dUnif}(x) = \int_0^1 f(u)\mathrm{d}u\,.$$

Now, assume that f satisfies (5) and let us see that it is a calibrator. Fix an arbitrary data space $\mathcal{X}$, a hypothesis $\mathcal{H}$ on $\mathcal{X}$, and a p-variable. For any fixed $Q \in \mathcal{H}$, we need to show that $\mathbb{E}_Q[f \circ p] \leq 1$. Since f is non-increasing and non-negative, replacing p by $\min\{p, 1\}$ can only increase $f(p)$, so it is enough to prove the bound under the additional assumption that p is bounded in $[0, 1]$. Under this assumption, we can define $F : [0, 1] \to [0, 1]$ as $F(t) = Q(p \leq t)$, and we have for all t

$$F(t) \leq t\,.$$

Let F^- denote the (generalised)[4] inverse of F. From $F(t) \leq t$ it follows that $F^-(u) \geq u$, for all $u \in [0, 1]$. Since F is the cdf of p, we have that p has the same law as $F^-(U)$, where U is uniformly distributed on $[0, 1]$.[5] So, we can write

$$\mathbb{E}_Q[f(p)] = \int_0^1 f\big(F^-(u)\big)\,du \leq \int_0^1 f(u)\mathrm{d}u \leq 1\,,$$

where we used that $f(F^-(u)) \leq f(u)$, since f is non-increasing. This shows that f is indeed a calibrator.

Example: testing a standard normal. Let us consider the case where

$$\mathcal{H} = \{Q\}\,, \qquad Q = \mathcal{N}(0, 1)^{\otimes n}\,.$$

So, a dataset is $X = (Y_1, \ldots, Y_n)$, with i.i.d. $Y_i \sim N(0, 1)$. Define $Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$, so that $Z \sim N(0, 1)$ under Q.

Let Φ denote the cdf of the standard normal. Then,

$$p : x \mapsto 1 - \Phi(z) = 1 - \Phi\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n y_i\right)$$

is a p-variable (uniform on $[0, 1]$ under Q), which tends to reject the null if large values of Z are observed. As p is uniformly distributed under the null, it is not a post-hoc e-variable (it is easily checked that $\mathbb{E}_Q[1/p] = \infty$).

It is easy to construct an e-variable, and hence a post-hoc p-variable. Using the fact that $\mathbb{E}_Q[e^{\lambda Z}] = e^{\lambda^2/2}$ (for any $\lambda \in \mathbb{R}$), we see that

$$E : x \mapsto e^{\sqrt{n}z - n/2} = \exp\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (y_i - 1/2)\right)$$

[4] $F^-(u) = \inf\{t \in [0, 1] : F(t) \geq u\}$.

[5] To see this, it is enough to show that $Y = F^-(U)$ has the same cdf as p. But this follows from the definition of F^- since $(F^-(u) \leq t) \iff (u \leq F(t))$, as F is non-decreasing.

is an e-variable. Hence, $1/E = e^{n/2-\sqrt{n}z}$ is a post-hoc p-variable.

Considering our previous discussion, a natural question is whether E can be obtained from p through a calibrator. Namely, if there is a calibrator f such that $E = f \circ p$. Solving for z in terms of p we see that $z = \Phi^{-1}(1-p)$, so if $E = f(p)$ we necessarily have

$$f : u \mapsto \exp\left(\sqrt{n}\Phi^{-1}(1-u) - n/2\right) .$$

Since $u \mapsto \Phi^{-1}(1-u)$ is strictly decreasing, f is non-increasing, and

$$\int_0^1 f(u)\mathrm{d}u = \mathbb{E}_{U\sim\mathrm{Unif}}[f(U)] = \mathbb{E}_Q[f(1-\Phi(Z))] = \mathbb{E}_Q\left[e^{\sqrt{n}Z-n/2}\right] = 1\,,$$

so f is a valid calibrator.

Note that not every e-variable can be obtained by p through the application of a calibrator! Indeed, the monotonicity requirement of the calibrator and the fact that p decreases with z imply that any $\tilde{E}$ in the form $\tilde{f}(p)$, for a calibrator $\tilde{f}$, must increase with z. Any e-variable that does not do so (e.g., the mapping $x \mapsto E(-x)$) cannot be generated by the specific p that we are considering through a calibrator. This simple example shows how a given p-variable generates via calibrators only those e-variables that preserve the appropriate monotone direction of evidence.

Bibliography. I have taken the idea to introduce e-variables as inverse of post-hoc p-variables from the paper *Post-hoc α Hypothesis Testing and the Post-hoc p-value* by Nick Koning. A closely related work, with a slightly different perspective, is *Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha* by Peter Grünwald. The discussion on calibrators is mostly based on Chapter 2 of the book *Hypothesis Testing with E-values* by Aaditya Ramdas and Ruodu Wang, which provides further context and related results.