E-Value Analysis of DESI Data: A Complete Walkthrough

Mathematics, Data, Processing, Testing, and Assumptions

Jinyoung Kim

February 2026

Abstract

This document walks through every step of our e-value analysis of DESI (Dark Energy Spectroscopic Instrument) data, from the underlying cosmological physics to the final statistical conclusions. For each step, we explain *why* we do it, *what* the math is, and give a concrete *worked example* with real numbers. The goal is that a reader with undergraduate-level mathematics can follow the entire chain: cosmological distances → BAO measurements → statistical framework (e-values) → results and their meaning.

Our central findings: (1) cross-dataset tension between DES-Y5 and Pantheon+ supernova catalogs is the most informative result (∼10,000× asymmetry in predictive e-values); (2) valid e-values from multiple methods converge on moderate evidence for w_0w_aCDM ($E \approx$ 10–20, or $\ln E \approx$ 2.3–3.0); (3) the data-split e-value ($E = 1.4$) is inconclusive due to limited statistical power for w_a, not evidence of absence. We are explicit about every assumption made and every limitation encountered.

Contents

Part I
The Physics

1 The Expanding Universe

The Core Idea

The universe is expanding. Galaxies are moving apart from each other, not because they are flying through space, but because space itself is stretching. The rate of this stretching depends on what the universe is made of.

1.1 Redshift

When a galaxy emits light at some time in the past, the light's wavelength gets stretched as the universe expands while the light travels to us. The **redshift** z quantifies this stretching:

$$1 + z = \frac{\lambda_{\text{observed}}}{\lambda_{\text{emitted}}} = \frac{1}{a} \tag{1}$$

where a is the **scale factor** of the universe at the time the light was emitted ($a = 1$ today).

Example: Redshift

A galaxy at $z = 1$ emitted its light when the universe was half its current size ($a = 1/2$). Light from a $z = 0.5$ galaxy was emitted when the universe was $2/3$ its current size. Higher z means further back in time, smaller universe.

1.2 The Hubble Parameter

The expansion rate is described by the **Hubble parameter**:

$$H(z) = H_0 \cdot E(z) \tag{2}$$

where $H_0 \approx 67.7$ km/s/Mpc is today's expansion rate and $E(z)$ is the **dimensionless Hubble parameter** that encodes how the expansion rate changes with redshift.

1.3 The Friedmann Equation

General relativity tells us $E(z)$ depends on the contents of the universe:

$$\boxed{E(z) = \sqrt{\Omega_m(1+z)^3 + \Omega_r(1+z)^4 + \Omega_k(1+z)^2 + \Omega_{DE}(z)}} \tag{3}$$

Each term represents a component of the universe:

Component	**What it is**	**Symbol**	**Value (Planck 2018)**
Matter	Dark matter + baryons	Ω_m	0.3111
Radiation	Photons + neutrinos	Ω_r	9×10^{-5}
Curvature	Spatial geometry	Ω_k	≈ 0
Dark energy	Accelerating expansion	Ω_{DE}	0.6889

Example: Computing $E(z)$ for ΛCDM

At $z = 0.7$ with the standard ΛCDM model (Ω_{DE} is constant):

$$E(0.7) = \sqrt{0.3111 \times 1.7^3 + 9 \times 10^{-5} \times 1.7^4 + 0 + 0.6889} \tag{4}$$

$$= \sqrt{0.3111 \times 4.913 + 0.0007 + 0.6889} \tag{5}$$

$$= \sqrt{1.528 + 0.0007 + 0.6889} \tag{6}$$

$$= \sqrt{2.218} = 1.489 \tag{7}$$

So at $z = 0.7$, the expansion rate is about 1.49 times today's rate.

Caution: Assumption: Flat Universe

We assume $\Omega_k = 0$ (spatially flat universe) throughout. This is well-supported by CMB data but is nonetheless an assumption. If the universe has slight curvature, distance calculations would change.

2 Dark Energy: The Central Question

2.1 The Cosmological Constant (ΛCDM)

In the simplest model, dark energy is Einstein's **cosmological constant** Λ with a fixed energy density. This is characterized by an **equation of state parameter** $w = -1$:

$$P_{DE} = w\rho_{DE}c^2 \quad \text{with } w = -1 \text{ (constant)} \tag{8}$$

Under ΛCDM, $\Omega_{DE}(z) = \Omega_{DE,0} = 0.6889$ — it does not change with redshift.

2.2 Dynamic Dark Energy (w_0w_aCDM)

What if dark energy is not constant? The CPL (Chevallier–Polarski–Linder) parametrization allows w to evolve:

$$\boxed{w(a) = w_0 + w_a(1 - a) = w_0 + w_a\frac{z}{1 + z}} \tag{9}$$

This gives two free parameters:

- w_0: the value of w today ($z = 0$, $a = 1$)
- w_a: how much w changes over time

The dark energy density then evolves as:

$$\Omega_{DE}(z) = \Omega_{DE,0} \cdot (1 + z)^{3(1+w_0+w_a)} \cdot \exp\left(-3w_a\frac{z}{1 + z}\right) \tag{10}$$

Example: DESI's Best-Fit Dynamic Dark Energy

DESI DR2 finds $w_0 \approx -0.75$, $w_a \approx -1.05$ as the best fit. Let's see what w looks like at different epochs:

z	$a = 1/(1+z)$	$w(a) = -0.75 + (-1.05)(1-a)$	Meaning
0	1.0	-0.75	Today: slightly less negative than -1
0.5	0.667	-1.10	More negative than Λ
1.0	0.5	-1.28	Even more negative
2.0	0.333	-1.45	Strongly phantom-like

Under this model, dark energy was *stronger* in the past ($w < -1$, "phantom") and is weakening toward $w \to -0.75$ today. ΛCDM has $w = -1$ at all times.

Caution: Assumption: CPL Parametrization

We assume dark energy dynamics (if any) can be captured by two numbers (w_0, w_a). This is a convenient but arbitrary choice. More complex evolution patterns would require different parametrizations.

3 Cosmological Distances

These are the quantities we actually compute and compare against data. All distances depend on $H(z)$, and therefore on the dark energy model.

3.1 Hubble Distance $D_H(z)$

The distance light would travel if the universe were expanding at the *instantaneous* rate at redshift z:

$$D_H(z) = \frac{c}{H(z)} = \frac{c}{H_0 E(z)} \tag{11}$$

This measures the expansion rate *at* redshift z directly.

Example: Hubble Distance

At $z = 0.7$ under ΛCDM (we computed $E(0.7) = 1.489$):

$$D_H(0.7) = \frac{299792.458}{67.66 \times 1.489} = \frac{299792.458}{100.73} = 2976 \text{ Mpc} \tag{12}$$

3.2 Comoving Distance $D_C(z)$ and Transverse Comoving Distance $D_M(z)$

The total comoving distance to redshift z is an integral over the whole line of sight:

$$D_C(z) = \frac{c}{H_0} \int_0^z \frac{dz'}{E(z')} \tag{13}$$

For a flat universe, $D_M = D_C$. This measures the total accumulated distance, sensitive to the expansion history at all redshifts between 0 and z.

3.3 Volume-Averaged Distance $D_V(z)$

For measurements that average over angles (isotropic BAO):

$$D_V(z) = \left[z \cdot D_H(z) \cdot D_M(z)^2\right]^{1/3} \tag{14}$$

This combines the radial (D_H) and transverse (D_M) distances.

3.4 The Sound Horizon r_d

Before the universe was 380,000 years old, photons and baryons formed a hot plasma with sound waves propagating through it. When the universe cooled enough for atoms to form ("recombination"), these waves froze in place. The distance the waves traveled is the **sound horizon**:

$$r_d = \int_{z_{\text{rec}}}^{\infty} \frac{c_s(z)}{H(z)}\, dz \approx 147 \text{ Mpc} \tag{15}$$

Why r_d Matters

$r_d \approx 147$ Mpc is a *known physical length* calibrated by well-understood early-universe physics. It acts as a "standard ruler." By measuring how big this ruler *appears* at different redshifts, we can infer distances. This is what BAO measures.

3.5 What We Actually Compare to Data

DESI measures distance *ratios* scaled by the sound horizon:

$$\frac{D_M(z)}{r_d}, \qquad \frac{D_H(z)}{r_d}, \qquad \frac{D_V(z)}{r_d} \tag{16}$$

These ratios are what our code computes (in `cosmology.py`) and compares to the measured values.

Example: Full Distance Calculation at $z = 0.706$

Under ΛCDM ($w_0 = -1, w_a = 0$), with $r_d = 147.09$ Mpc:
$D_H(0.706)/r_d$: We need $E(0.706) \approx 1.497$, so:

$$\frac{D_H}{r_d} = \frac{c/(H_0 \cdot E(z))}{r_d} = \frac{299792.458/(67.66 \times 1.497)}{147.09} = \frac{2960}{147.09} \approx 20.12$$

$D_M(0.706)/r_d$: Requires numerical integration $\int_0^{0.706} dz'/E(z')$. The result is $D_M/r_d \approx 17.86$.
The DESI DR2 measurements at $z = 0.706$ are: $D_M/r_d = 17.35 \pm 0.18$, $D_H/r_d = 19.46 \pm 0.33$.

Part II
The Data

4 DESI DR2 BAO Measurements

Data: Source and Format

Source: Official DESI public data release via `CobayaSampler/bao_data` on GitHub.
Files:

- `desi_gaussian_bao_ALL_GCcomb_mean.txt`: 13 measurements
- `desi_gaussian_bao_ALL_GCcomb_cov.txt`: 13 × 13 covariance matrix

4.1 The Data Vector

The complete DESI DR2 dataset consists of 13 BAO measurements at 7 effective redshifts:

Index	z_{eff}	Quantity	Value	Tracer
1	0.295	D_V/r_d	7.942	BGS
2	0.510	D_M/r_d	13.588	LRG1
3	0.510	D_H/r_d	21.863	LRG1
4	0.706	D_M/r_d	17.351	LRG2
5	0.706	D_H/r_d	19.455	LRG2
6	0.934	D_M/r_d	21.576	LRG3+ELG1
7	0.934	D_H/r_d	17.641	LRG3+ELG1
8	1.321	D_M/r_d	27.601	ELG2
9	1.321	D_H/r_d	14.176	ELG2
10	1.484	D_M/r_d	30.512	QSO
11	1.484	D_H/r_d	12.817	QSO
12	2.330	D_H/r_d	8.632	Lyα
13	2.330	D_M/r_d	38.989	Lyα

4.2 Understanding the Tracers

DESI uses different types of objects to measure BAO at different redshifts:

Tracer	Redshift Range	What it is
BGS	$z \sim 0.3$	Bright Galaxy Survey: nearby bright galaxies
LRG	$z \sim 0.5$–0.9	Luminous Red Galaxies: massive, red galaxies
ELG	$z \sim 0.9$–1.3	Emission Line Galaxies: star-forming galaxies
QSO	$z \sim 1.5$	Quasars: active galactic nuclei
Lyα	$z \sim 2.3$	Lyman-α forest: absorption in quasar spectra

4.3 The Covariance Matrix

The 13×13 covariance matrix C encodes both measurement uncertainties and correlations. It is **block-diagonal**: measurements at different redshifts are uncorrelated, but D_M/r_d and D_H/r_d at the *same* redshift are correlated (typically anti-correlated).

Example: Reading the Covariance Matrix

At $z = 0.706$ (indices 4 and 5 in the data vector), the covariance block is:

$$C_{z=0.706} = \begin{pmatrix} 0.0324 & -0.0237 \\ -0.0237 & 0.1115 \end{pmatrix}$$

This tells us:

- $\sigma(D_M/r_d) = \sqrt{0.0324} = 0.180$ (1.0% precision)
- $\sigma(D_H/r_d) = \sqrt{0.1115} = 0.334$ (1.7% precision)
- Correlation: $\rho = \frac{-0.0237}{\sqrt{0.0324 \times 0.1115}} = -0.39$ (anti-correlated)

The anti-correlation is physical: if the BAO peak is measured to be at a slightly larger angle (larger D_M), the corresponding radial measurement (D_H) tends to be slightly smaller.

Caution: Assumption: Gaussian Errors

We assume the likelihood is multivariate Gaussian: $\mathcal{L}(\mathbf{d}|\boldsymbol{\theta}) \propto \exp\big(-\frac{1}{2}\chi^2\big)$. This is standard for BAO summary statistics and validated by DESI, but it is an approximation. Non-Gaussianity could affect tail probabilities.

Part III
The Statistical Framework

5 Hypothesis Testing: The Setup

What We Are Testing

- H_0 (**Null**): The universe has a cosmological constant. $w_0 = -1$, $w_a = 0$.
- H_1 (**Alternative**): Dark energy is dynamic. w_0 and w_a are free parameters.

We want to know: does the DESI data provide compelling evidence that H_0 is wrong and H_1 is better?

5.1 The χ^2 Statistic

The fundamental measure of fit quality is the chi-squared statistic:

$$\boxed{\chi^2 = (\mathbf{d} - \mathbf{t})^T \, C^{-1} \, (\mathbf{d} - \mathbf{t})} \tag{17}$$

where:

- $\mathbf{d}$ is the data vector (13 measurements)
- $\mathbf{t}(\boldsymbol{\theta})$ is the theory prediction vector (depends on cosmological parameters)
- C is the 13×13 covariance matrix
- C^{-1} is its inverse

Smaller χ^2 means better fit. The improvement of H_1 over H_0 is:

$$\Delta\chi^2 = \chi^2_{H_0} - \chi^2_{H_1} \tag{18}$$

Example: χ^2 With Real Numbers

For DESI DR2, the results are:

$$\chi^2_{\Lambda\text{CDM}} \approx 25.4 \quad \text{(13 data points, 0 free params)} \tag{19}$$

$$\chi^2_{w_0 w_a\text{CDM}} \approx 13.5 \quad \text{(13 data points, 2 free params)} \tag{20}$$

$$\Delta\chi^2 = 25.4 - 13.5 = 11.9 \tag{21}$$

With 2 extra parameters and $\Delta\chi^2 = 11.9$, the proper frequentist calculation uses the $\chi^2(2)$ distribution: $p = 1 - F_{\chi^2}(11.9; 2) = e^{-5.95} \approx 0.0026$, which corresponds to $\approx 2.8\sigma$ (two-sided). Note: the sometimes-quoted 3.5σ uses $\sqrt{\Delta\chi^2}$, which is only valid for $k = 1$ extra parameter (see Section 11.3).

But is this real evidence, or just overfitting? This is where e-values come in.

5.2 The Log-Likelihood

Under Gaussian errors, the log-likelihood is:

$$\ln \mathcal{L} = -\frac{1}{2}\chi^2 - \frac{1}{2}\ln|C| - \frac{n}{2}\ln(2\pi) \tag{22}$$

The last two terms are constants (they don't depend on the model), so differences in log-likelihood are determined entirely by $\Delta\chi^2$:

$$\ln \mathcal{L}_{H_1} - \ln \mathcal{L}_{H_0} = -\frac{1}{2}(\chi^2_{H_1} - \chi^2_{H_0}) = \frac{\Delta\chi^2}{2} \tag{23}$$

6 E-Values: What They Are and Why We Use Them

6.1 The Problem with Just Using $\Delta\chi^2$

Why not just report $\Delta\chi^2 = 11.9$ and declare victory?

Caution: The Overfitting Problem

When we fit (w_0, w_a) to the same data we use to evaluate $\Delta\chi^2$, the improvement is *guaranteed* to be positive even if H_0 is true. Two free parameters can always improve the fit by chance. This inflates $\Delta\chi^2$ and makes the evidence look stronger than it really is.

A correction factor exists (Wilks' theorem says $\Delta\chi^2$ should follow a χ^2 distribution with 2 degrees of freedom under H_0), but this relies on regularity conditions that may not hold here, and it doesn't address whether the *specific* parameter values generalize.

6.2 Definition of an E-Value

Definition 1 (E-Value). *A non-negative random variable E is an* ***e-value*** *for testing H_0 if:*

$$\boxed{\mathbb{E}[E \mid H_0] \leq 1} \tag{24}$$

That is, the expected value of E under the null hypothesis is at most 1.

Intuition

Think of E as "evidence multiplied." Under H_0:

- On average, $E \leq 1$ (you don't expect to accumulate false evidence)
- A large E is surprising — it's unlikely under H_0
- By Markov's inequality: $\mathbb{P}_{H_0}(E \geq 1/\alpha) \leq \alpha$

So if $E = 100$, the probability of seeing $E \geq 100$ under H_0 is at most $1/100 = 1\%$.

6.3 E-Values vs. P-Values

	P-value	**E-value**
Definition	$P(\text{more extreme data} \mid H_0)$	Random variable with $\mathbb{E}[E \mid H_0] \leq 1$
Interpretation	How surprising is the data?	How much evidence against H_0?
Combining	Complex (Fisher's method, etc.)	Simple: multiply (if independent)
Optional stopping	Invalidates the test	Still valid
Overfitting risk	High if model is fitted post-hoc	Can be controlled (data-splitting, mixtures)

6.4 Interpreting E-Values

E-value	**Interpretation**
$E < 1$	Evidence *favors* H_0 (null)
$E = 1$	No evidence either way
$E \sim 3$	Weak evidence against H_0
$E \sim 10$	Moderate evidence against H_0
$E \sim 100$	Strong evidence against H_0
$E \sim 1000$	Very strong evidence

Approximate sigma conversion: For rough comparison with frequentist significance, one can use $\sigma \approx \sqrt{2 \ln E}$. For example, $E = 100$ gives $\sigma \approx 3.0$. But this conversion is approximate; e-values and p-values answer different questions.

7 Four Methods to Compute E-Values

We implement four methods, each with different validity properties and trade-offs.

7.1 Method 1: Maximized Likelihood Ratio (Not a Valid E-Value)

The Idea

The most basic approach: how much better does H_1 fit the data compared to H_0? This produces a likelihood ratio statistic, but **not** a valid e-value when the alternative is fitted to the same data (see below).

$$E = \frac{\mathcal{L}(\mathbf{d} \mid H_1)}{\mathcal{L}(\mathbf{d} \mid H_0)} = \exp\left(\frac{\Delta\chi^2}{2}\right) \tag{25}$$

When this is a valid e-value (for *fixed* H_1):
If H_1 is fully specified before seeing the data, then under H_0:

$$\mathbb{E}_{H_0}[E] = \int \frac{P_1(x)}{P_0(x)} P_0(x)\, dx = \int P_1(x)\, dx = 1 \tag{26}$$

Why plugging in the MLE is not an e-value:

If we first fit (w_0, w_a) to the data and *then* compute this ratio, the proof above does not apply. Worse, the expectation diverges. Under H_0, the $\Delta\chi^2$ for $k = 2$ extra parameters follows a $\chi^2(2)$ distribution, so:

$$\mathbb{E}_{H_0}\left[\exp\left(\tfrac{\chi^2(2)}{2}\right)\right] = \int_0^\infty e^{t/2} \cdot \tfrac{1}{2}e^{-t/2}\,dt = \int_0^\infty \tfrac{1}{2}\,dt = \infty \tag{27}$$

Since $\mathbb{E}[E \mid H_0] = \infty$, the defining property $\mathbb{E}[E \mid H_0] \leq 1$ is violated. **The maximized likelihood ratio is not an e-value**—it is a descriptive statistic only.

Example: Maximized Likelihood Ratio (NOT AN E-VALUE)

With DESI DR2 and the DESI best-fit $w_0 = -0.75$, $w_a = -1.05$:

$$\Delta\chi^2 \approx 11.9 \tag{28}$$

$$\mathrm{LR}_{\max} = e^{11.9/2} = e^{5.95} \approx 384 \tag{29}$$

Looks very strong! But this is **not an e-value** ($\mathbb{E}[E|H_0] = \infty$). The alternative was fitted to the same data used for evaluation, producing a maximized likelihood ratio that cannot be interpreted as calibrated evidence.

7.2 Method 2: Uniform Mixture E-Value

The Idea

Instead of using one specific (w_0, w_a), average the likelihood ratio over a grid of possible alternatives with uniform weights. This "pre-specifies" the alternative as a mixture, preventing cherry-picking. We use a uniform mixture over a grid of alternatives (sometimes called a "mixture e-value"). This is related to but distinct from the GROW-optimal procedure of Grünwald, de Heide & Koolen (2024), which would optimize the mixture weights to maximize the expected log growth rate.

$$E_{\mathrm{mix}} = \int \frac{\mathcal{L}(\mathbf{d} \mid w_0, w_a)}{\mathcal{L}(\mathbf{d} \mid H_0)}\,\pi(w_0, w_a)\,dw_0\,dw_a \tag{30}$$

In practice, we discretize: define a 10×10 grid over $(w_0, w_a) \in [-1.5, -0.5] \times [-2.0, 1.0]$, compute the likelihood ratio at each grid point, and average (with uniform weights):

$$E_{\mathrm{mix}} = \frac{1}{100}\sum_{i=1}^{100} \exp\left(\frac{\chi^2_{H_0} - \chi^2_i}{2}\right) \tag{31}$$

Why this is valid: Each grid point gives a valid e-value (since H_1 is specified before seeing data). The average of e-values is an e-value (by linearity of expectation).

The trade-off: prior sensitivity. The result depends on which grid range we use:

Prior Range	w_0 range	w_a range	**E-value**
Narrow	$[-1.2, -0.8]$	$[-1.0, 0.5]$	~ 97
Default	$[-1.5, -0.5]$	$[-2.0, 1.0]$	~ 15
Wide	$[-2.0, 0.0]$	$[-3.0, 2.0]$	~ 17

The variation from 15 to 97 reflects the Occam razor operating as expected: a narrower prior concentrating mass near the MLE wastes less probability on distant parameter values and therefore produces a larger e-value. This is a standard feature of Bayesian and mixture-based tests, not a deficiency. The default range $E \approx 15$ constitutes moderate-to-strong evidence ($\ln E \approx 2.7$).

7.3 Method 3: Data-Split E-Value

The Idea

Split the data into two parts. Use one part to train (fit the alternative), and the other part to test. Because the test data was never used for fitting, the resulting e-value is honest.

Procedure:

1. **Split**: Divide 13 measurements into training set (7 measurements at $z < 1$) and test set (6 measurements at $z \geq 1$).
2. **Train**: Fit (w_0, w_a) by minimizing χ^2 on the training data only.
3. **Test**: Compute the likelihood ratio on the test data using the fitted parameters.

$$\boxed{E_{\text{split}} = \frac{\mathcal{L}(D_{\text{test}} \mid \hat{w}_0, \hat{w}_a)}{\mathcal{L}(D_{\text{test}} \mid H_0)} = \exp\left(\frac{\chi^2_{\text{test},H_0} - \chi^2_{\text{test},H_1}}{2}\right)} \tag{32}$$

Why this is valid:

Conditional on D_{train}, the fitted parameters $(\hat{w}_0, \hat{w}_a)$ are fixed constants. From D_{test}'s perspective, H_1 is fully specified before the test data is observed. Therefore:

$$\mathbb{E}[E_{\text{split}} \mid D_{\text{train}}] = 1 \quad \Rightarrow \quad \mathbb{E}[E_{\text{split}}] = 1 \tag{33}$$

Example: Data-Split E-Value (VALID but UNDERPOWERED)

Step 1: Split the data.
Training set ($z < 1$): 7 measurements at $z = 0.295, 0.51, 0.51, 0.706, 0.706, 0.934, 0.934$.
Test set ($z \geq 1$): 6 measurements at $z = 1.321, 1.321, 1.484, 1.484, 2.33, 2.33$.
Step 2: Extract sub-covariance matrices.
C_{train} is the 7×7 submatrix (rows/cols 1–7 of the full matrix).
C_{test} is the 6×6 submatrix (rows/cols 8–13). Because the full covariance is block-diagonal across redshift bins, these blocks are independent.
Step 3: Fit on training data.
Minimize $\chi^2_{\text{train}}(w_0, w_a)$ using L-BFGS-B optimizer with bounds $w_0 \in [-2, 0]$, $w_a \in [-3, 2]$. Starting point: $w_0 = -0.9$, $w_a = -0.5$.
Result: $\hat{w}_0 = -0.78$, $\hat{w}_a = -0.52$.
Step 4: Evaluate on test data.
Compute χ^2_{test} under both models:

$$\chi^2_{\text{test, }\Lambda\text{CDM}} \approx 5.8 \tag{34}$$

$$\chi^2_{\text{test, }w_0 w_a} \approx 5.1 \tag{35}$$

$$\Delta\chi^2_{\text{test}} = 0.7 \tag{36}$$

Step 5: Compute e-value.

$$E_{\text{split}} = e^{0.7/2} = e^{0.35} \approx 1.4 \tag{37}$$

Interpretation: The alternative model predicts high-redshift data only 1.4 times better than ΛCDM. This is essentially no evidence — but see the critical power caveat below.

Caution: The Data-Split Power Problem for w_a

The data-split e-value of $E = 1.4$ is **inconclusive**, not evidence against w_0w_aCDM. The key issue is that the redshift-based split creates a severe power problem for w_a:
In the CPL parameterization $w(a) = w_0 + w_a(1 - a)$, the w_a contribution is $(1 - a) \cdot w_a$. For the training set at $z < 1$:

- At $z = 0.295$: $a = 0.77$, so $(1 - a) = 0.23$ — only 23% leverage on w_a
- At $z = 0.934$: $a = 0.52$, so $(1 - a) = 0.48$ — only 48% leverage on w_a

The training set has limited ability to constrain w_a, so the fitted parameters $(\hat{w}_0, \hat{w}_a)$ may be far from the true values.
Power calibration confirms this: Monte Carlo simulations (500 realizations) show that *even when w_0w_aCDM is the true model generating the data*, the data-split test yields a median $E_{\text{split}} \approx 2.7$. The observed $E = 1.4$ is fully consistent with the signal being real but the test being underpowered.

Caution: Assumption: Independence of Training and Test Sets

The validity of data-split e-values requires D_{train} and D_{test} to be independent. This holds here because the DESI covariance matrix is block-diagonal across redshift bins: measurements at different redshifts are uncorrelated. If there were cross-redshift correlations (e.g., from systematic effects), this assumption would be violated.

7.4 Method 4: Leave-One-Out (LOO) Average E-Value

The Idea

Instead of a single train/test split, we leave out one redshift bin at a time, fit on the remaining bins, and compute an e-value for the held-out bin. We then **average** (not multiply) the per-bin e-values. This uses data more efficiently than a single split.

Procedure: For each of the $K = 7$ redshift bins, indexed by k:

1. Remove all measurements at redshift bin k (1 or 2 data points).
2. Fit (w_0, w_a) on the remaining data, obtaining $\hat{\theta}_{-k}$.
3. Compute the e-value for the held-out bin:

$$E_k = \frac{\mathcal{L}(D_k \mid \hat{\theta}_{-k})}{\mathcal{L}(D_k \mid H_0)} \tag{38}$$

The **LOO average** is:

$$\boxed{E_{\text{LOO-avg}} = \frac{1}{K}\sum_{k=1}^{K} E_k} \tag{39}$$

Why the average is valid: Each individual E_k is a valid e-value (conditional on the training data, the held-out data tests a pre-specified alternative). The average of e-values is an e-value by linearity of expectation:

$$\mathbb{E}\left[\frac{1}{K}\sum_k E_k \,\middle|\, H_0\right] = \frac{1}{K}\sum_k \mathbb{E}[E_k \mid H_0] \leq \frac{1}{K}\cdot K = 1 \tag{40}$$

Caution: Why the LOO *product* is NOT valid

It is tempting to multiply the LOO e-values ($\prod_k E_k$), since products of independent e-values are e-values. However, the LOO e-values are **not independent**: the training sets overlap. For example, E_1 uses $\hat{\theta}_{-1}$ (fitted on bins 2–7) and E_2 uses $\hat{\theta}_{-2}$ (fitted on bins 1, 3–7). Both training sets contain bin 3, so E_1 and E_2 are dependent through shared training data.

The product $\prod_k E_k$ is NOT guaranteed to satisfy $\mathbb{E}[\prod E_k \mid H_0] \leq 1$. The mutual dependence can inflate the expectation above 1, invalidating the e-value property.

Example: LOO Average E-Value

The per-bin LOO e-values for DESI DR2:

Left-out bin	z_{eff}	$(\hat{w}_0, \hat{w}_a)_{-k}$	E_k
BGS	0.295	$(-0.86, -0.44)$	1.89
LRG1	0.510	$(-0.84, -0.46)$	0.71
LRG2	0.706	$(-0.85, -0.38)$	55.98
LRG3+ELG1	0.934	$(-0.84, -0.49)$	8.73
ELG2	1.321	$(-0.86, -0.42)$	2.14
QSO	1.484	$(-0.86, -0.43)$	0.75
Lyα	2.330	$(-0.86, -0.44)$	1.00

LOO average:

$$E_{\text{LOO-avg}} = \frac{1.89 + 0.71 + 55.98 + 8.73 + 2.14 + 0.75 + 1.00}{7} = \frac{71.2}{7} \approx 10.2$$

The evidence is concentrated at $z = 0.706$ (LRG2, $E_k = 56$) and $z = 0.934$ (LRG3+ELG1, $E_k = 8.7$). This is consistent with the signal being strongest in the intermediate-redshift range where the BAO measurements are most precise. The fitted parameters are stable across folds ($w_0 \approx -0.85$, $w_a \approx -0.4$), suggesting a consistent signal rather than noise.

The LOO product $\prod_k E_k = 1.89 \times 0.71 \times 55.98 \times \ldots \approx 1062$ would appear very strong. But this is **not valid** due to overlapping training sets (see caution above).

Part IV
The Processing Pipeline

8 Step-by-Step: What the Code Does

Here is the complete processing chain, corresponding to the code modules:

8.1 Step 1: Load Data (`data_loader.py`)

1. Read `desi_gaussian_bao_ALL_GCcomb_mean.txt`: parse each line as $(z, \text{value}, \text{quantity})$.

2. Read `desi_gaussian_bao_ALL_GCcomb_cov.txt`: load 13 × 13 matrix.
3. Infer tracer type from redshift (e.g., $z = 0.295 \rightarrow$ BGS).
4. Package into a `BAODataset` object with arrays for z, data, covariance, quantities, tracers.

Assumption: The files are unmodified from the official release.

8.2 Step 2: Compute Theory Predictions (`cosmology.py`)

For given cosmological parameters (w_0, w_a):

1. Compute $E(z)$ using Eq. (3) with dark energy evolution from Eq. (10).
2. Compute $D_H(z) = c/[H_0 E(z)]$.
3. Compute $D_C(z) = \frac{c}{H_0} \int_0^z dz'/E(z')$ via numerical integration (`scipy.integrate.quad`).
4. Compute $D_M(z) = D_C(z)$ (flat universe).
5. Compute $D_V(z) = [z \cdot D_H \cdot D_M^2]^{1/3}$.
6. Divide by $r_d = 147.09$ Mpc.
7. Build theory vector matching the order of the data vector (which entries are D_M/r_d, which are D_H/r_d, which are D_V/r_d).

8.3 Step 3: Compute χ^2 (`cosmology.py`)

$$\chi^2 = (\mathbf{d} - \mathbf{t})^T C^{-1} (\mathbf{d} - \mathbf{t}) \tag{41}$$

Implemented as: compute residual vector $\mathbf{r} = \mathbf{d} - \mathbf{t}$, invert covariance via `numpy.linalg.inv`, compute quadratic form.

Caution: Assumption: Invertible Covariance

We compute C^{-1} directly. This works because C is 13×13 and well-conditioned (condition number $\sim 10^3$). For larger matrices, one would use Cholesky decomposition for numerical stability.

8.4 Step 4: Fit the Alternative Model (`evalue_analysis.py`)

For the data-split and LOO methods:

1. Extract training data subset and sub-covariance matrix.
2. Define objective: $\chi^2_{\text{train}}(w_0, w_a)$.
3. Minimize using L-BFGS-B with bounds $w_0 \in [-2, 0]$, $w_a \in [-3, 2]$.
4. Output: best-fit $\hat{w}_0, \hat{w}_a$.

8.5 Step 5: Compute E-Value (`evalue_analysis.py`)

For each method:

1. Compute theory vectors under H_0 and H_1.
2. Compute χ^2 under both models.
3. Compute $E = \exp(\Delta\chi^2/2)$.

Part V
The Testing and Results

9 Main Results

9.1 Cross-Dataset E-Values: The Strongest Finding

We begin with what we consider the most informative result. Cross-dataset e-values test whether $w_0 w_a$CDM parameters fitted on one experiment predict another experiment better than ΛCDM. Unlike within-dataset tests, this is not vulnerable to the power objection: if two experiments both detect the same underlying physics, parameters from one *must* predict the other well.

Train on	**Test on**	(w_0, w_a)	**E-value**	**Interpretation**
DESI (fitted)	Pantheon+	$(-0.86, -0.43)$	1.5	No evidence
DESI (fitted)	DES-Y5	$(-0.86, -0.43)$	86	Moderate
Pantheon+	DESI	$(-0.90, -0.20)$	2049	Strong
DES-Y5	DESI	$(-0.65, -1.20)$	**0.19**	*Favors ΛCDM!*

The ∼10,000× Asymmetry

The central finding: DES-Y5's best-fit $w_0 w_a$CDM parameters predict DESI data *worse* than ΛCDM ($E = 0.19 < 1$), while Pantheon+'s parameters predict DESI data well ($E = 2049$). This ∼10,000× asymmetry means the two leading supernova catalogs disagree on what dark energy dynamics look like.

If $w_0 w_a$CDM represents real physics, all experiments should measure the same equation of state and therefore make compatible predictions. The fact that DES-Y5's preferred parameters ($w_0 = -0.65$, $w_a = -1.20$) are actively *harmful* for predicting DESI data means these datasets are pulling in incompatible directions. This raises the possibility that $w_0 w_a$CDM is absorbing inter-dataset tension rather than detecting new physics.

9.2 Valid Within-Dataset E-Values

The valid e-value methods converge on moderate evidence:

Method	**E-value**	$\ln E$	$\sim\sigma$	**Valid?**
Uniform mixture (narrow)	97	4.57	3.0	Yes (prior-sensitive)
Uniform mixture (default)	**15**	**2.71**	**2.3**	**Yes**
Uniform mixture (wide)	17	2.83	2.4	Yes (prior-sensitive)
LOO average	**10.2**	**2.32**	**2.2**	**Yes**
Data-split ($z = 1$)	1.4	0.34	0.8	Yes, but underpowered
Maximized LR	392	5.97	—	**NOT AN E-VALUE**
LOO product	1062	6.97	—	**NOT VALID**

Notes:

- The $\sim\sigma$ column uses the approximate conversion $\sigma \approx \sqrt{2\ln E}$.
- The maximized LR has $\mathbb{E}[E|H_0] = \infty$ (not an e-value at all; see Method 1).
- The LOO product has $\mathbb{E}[\prod E_k|H_0]$ unknown (overlapping training sets; see Section 7.4).
- No sigma equivalent is shown for invalid statistics.

Where the Valid Methods Converge

The two independent valid methods — uniform mixture and LOO average — give consistent results:

$$E_{\text{mixture}} \approx 15 \quad (\ln E = 2.7) \tag{42}$$

$$E_{\text{LOO-avg}} \approx 10 \quad (\ln E = 2.3) \tag{43}$$

Both indicate moderate evidence for w_0w_aCDM over ΛCDM within DESI BAO data alone. The data-split $E = 1.4$ is consistent with these values once power loss is accounted for (median $E_{\text{split}} \approx 2.7$ under H_1).

9.3 Data-Split Power Calibration

To understand whether $E = 1.4$ is evidence against w_0w_aCDM or simply reflects low power, we performed Monte Carlo power calibration: generating 500 synthetic datasets under the w_0w_aCDM model (the *alternative* hypothesis is true) and computing E_{split} for each.

Quantity	**Value**
Median E_{split} under H_1 (signal is real)	2.7
$P(E_{\text{split}} > 1.4 \mid H_1)$	64%
$P(E_{\text{split}} > 1.4 \mid H_0)$	12.8%

Interpretation: Even when w_0w_aCDM is the true model, the data-split test typically gives only $E \approx 2.7$. The observed $E = 1.4$ cannot distinguish between H_0 and H_1. The data-split result is **inconclusive**, not negative.

9.4 Cross-Dataset Validation

We also test whether parameters from other experiments predict DESI data better than ΛCDM. See Section 9.1 above.

10 DR1 → DR2 Temporal Validation

DESI released DR1 (Year 1, 2024) and DR2 (Years 1–3, 2025). We test: do parameters fitted on DR1 predict DR2?

Scenario	$\Delta\chi^2$	**E-value**	**Note**
DR2 fit on DR2	∼12	∼400	NOT AN E-VALUE
DR1 fit predicting DR2	varies	varies	Semi-valid*

***Caveat**: DR2 *contains* DR1 (DR2 is 3 years of data including Year 1). They are not independent, so this is a test of *stability* rather than true out-of-sample prediction.

11 Information Criteria and the Sigma Conversion

In addition to e-values, there are other standard ways to assess the evidence for w_0w_aCDM over ΛCDM. We present information criteria and clarify the proper frequentist σ-conversion.

11.1 AIC and BIC from $\Delta\chi^2$

Both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) start from the χ^2 statistic and add a penalty for model complexity.

Information Criteria Formulas

For model comparison, the relevant quantities are:

$$\text{AIC} = \chi^2 + 2k \tag{44}$$

$$\text{BIC} = \chi^2 + k \ln n \tag{45}$$

where k is the number of free parameters and n is the number of data points. Lower values indicate a better model. For comparing two models:

$$\Delta\text{AIC} = \Delta\chi^2 - 2\Delta k \tag{46}$$

$$\Delta\text{BIC} = \Delta\chi^2 - \Delta k \cdot \ln n \tag{47}$$

where $\Delta k = 2$ (the extra parameters in w_0w_aCDM) and positive Δ favors the more complex model.

Example: Computing ΔAIC and ΔBIC

From our DESI analysis: $\Delta\chi^2 = 11.9$, $\Delta k = 2$, $n = 13$ data points.
AIC:

$$\Delta\text{AIC} = 11.9 - 2 \times 2 = 11.9 - 4 = 7.9 \tag{48}$$

By the Burnham & Anderson (2002) scale: $\Delta\text{AIC} > 6$ is “strong” evidence. So AIC **strongly favors w_0w_aCDM**.
BIC:

$$\Delta\text{BIC} = 11.9 - 2 \times \ln(13) = 11.9 - 2 \times 2.565 = 11.9 - 5.13 = 6.8 \tag{49}$$

By the Kass & Raftery (1995) scale: $6 < \Delta\text{BIC} < 10$ is “strong” evidence. So BIC also **favors w_0w_aCDM**.

Caution: The BIC Sample Size Problem

In the BIC formula, what is n? We used $n = 13$ (the number of BAO summary statistics). But DESI observed millions of galaxy pairs to produce these 13 summary statistics. If we used $n = 10^6$:

$$\Delta\text{BIC} = 11.9 - 2 \times \ln(10^6) = 11.9 - 27.6 = -15.7$$

This would **strongly favor ΛCDM**! The ambiguity of n in cosmological model selection is a well-known problem discussed by Liddle (2007). There is no consensus on the correct choice.

11.2 Why AIC/BIC and the Bayes Factor Disagree

This is an important point: AIC and BIC both favor w_0w_aCDM, yet the Bayes factor (from Ong et al. 2025) favors ΛCDM. How is this possible?

Different Complexity Penalties

The key difference is how each method penalizes model complexity:

Method	Penalty	Nature
AIC	$2k = 4$	Fixed, small
BIC ($n = 13$)	$k \ln n \approx 5.1$	Fixed, moderate
Bayes factor	Full prior volume	Depends on prior width

AIC has *no* Occam factor — it penalizes only by the number of parameters, regardless of how wide the prior range is. BIC *approximately* includes an Occam factor through the $k \ln n$ term. The Bayes factor includes the *full* Occam penalty: with broad priors over (w_0, w_a), much of the prior volume falls on parameter values that fit the data poorly. This "wasted" prior mass penalizes the complex model.

The uniform mixture e-value ($E \approx 15$) sits between AIC/BIC and the Bayes factor, because it averages over a grid of alternatives — a form of prior integration, but with a restricted range.

11.3 The Correct Frequentist σ-Conversion

The conversion from $\Delta\chi^2$ to "number of sigma" is often done incorrectly. Here is the proper procedure.

σ-Conversion for k Extra Parameters

Under H_0, the statistic $\Delta\chi^2$ follows a χ^2 distribution with k degrees of freedom (by Wilks' theorem). The p-value is:

$$p = 1 - F_{\chi^2}(\Delta\chi^2;\, k) \tag{50}$$

For $k = 2$, the CDF has a simple closed form: $F_{\chi^2}(x;\, 2) = 1 - e^{-x/2}$, so:

$$p = e^{-\Delta\chi^2/2} \tag{51}$$

The equivalent Gaussian significance (two-sided) is $\sigma = \Phi^{-1}(1 - p/2)$.

Example: σ-Conversion for DESI's $\Delta\chi^2 = 11.9$

Wrong method ($k = 1$ formula): $\sigma = \sqrt{\Delta\chi^2} = \sqrt{11.9} \approx 3.45\sigma$.
This is only valid when there is 1 extra parameter. It treats $\Delta\chi^2$ as if it were the square of a single standard normal variable.
Correct method ($k = 2$):

$$p = e^{-11.9/2} = e^{-5.95} \approx 0.0026 \tag{52}$$

$$\sigma = \Phi^{-1}(1 - 0.0026/2) = \Phi^{-1}(0.9987) \approx 2.8\sigma \tag{53}$$

The difference is significant: 2.8σ vs. 3.5σ. The $k = 1$ formula overestimates the significance by about 0.7σ because it does not account for the fact that with 2 free parameters, a large $\Delta\chi^2$ is more likely to occur by chance.

For reference, a few values:

$\Delta\chi^2$	σ ($k=1$, **wrong**)	σ ($k=2$, **correct**)
6.0	2.4	1.8
9.2	3.0	2.3
11.9	3.5	2.8
13.8	3.7	3.0

11.4 The Full Evidence Landscape

Putting it all together, here is the complete picture of evidence for w_0w_aCDM vs. ΛCDM:

Method	**Value**	**Favors**	**Validity**
Frequentist p ($k=2$)	$p=0.0026$ (2.8σ)	w_0w_aCDM	Same-data (no Occam)
ΔAIC	7.9	w_0w_aCDM	Penalty: $2k=4$
ΔBIC ($n=13$)	6.8	w_0w_aCDM	Penalty: $k\ln n\approx 5$
Uniform mixture E-value	$E\approx 15$	w_0w_aCDM	Valid e-value
LOO average E-value	$E\approx 10$	w_0w_aCDM	Valid e-value
Bayes factor (Ong et al.)	$\ln\mathcal{B}=-0.57$	ΛCDM	Full prior volume
Data-split E-value	$E=1.4$	Inconclusive	Valid, but underpowered
Cross-dataset (DES-Y5→DESI)	$E=0.19$	Tension	Valid

The pattern is clear: methods with *smaller* complexity penalties favor w_0w_aCDM, while the Bayes factor with its full prior penalty favors ΛCDM. The valid e-values from two independent methods (mixture and LOO) converge on $E\approx$ 10–15 (moderate evidence). The data-split is inconclusive due to power loss. And the cross-dataset tension ($E=0.19$) raises the question of whether the signal reflects real physics or dataset inconsistency.

Part VI
Assumptions, Limitations, and Caveats

12 Complete List of Assumptions

We organize every assumption into categories with an assessment of how critical each one is.

12.1 Cosmological Assumptions

1. **Flat universe** ($\Omega_k=0$). *Impact: Low.* Well-supported by CMB.

2. **CPL parametrization** for dark energy ($w=w_0+w_a(1-a)$). *Impact: Medium.* A different parametrization could change results.

3. **Planck 2018 fiducial parameters** ($\Omega_m=0.3111$, $h=0.6766$, $r_d=147.09$ Mpc). *Impact: Low.* Small changes in these don't qualitatively change results.

4. **Standard distance formulas** without full Boltzmann code (CAMB/CLASS). *Impact: Low–Medium.* Our distances agree with DESI's to $<0.5\%$.

12.2 Statistical Assumptions

5. **Gaussian likelihood**. *Impact: Low.* Standard for BAO summary statistics; validated by DESI.

6. **Published covariance matrix is correct**. *Impact: Medium.* We rely entirely on DESI's error estimates.

7. **Independence of redshift bins** (block-diagonal covariance). *Impact: Medium.* This is critical for the data-split and LOO e-value validity. Cross-redshift systematics would violate this.

8. **Data-split and LOO e-value validity** (training and test sets are independent). *Impact: High.* Follows from assumption 7.

12.3 Methodological Assumptions

9. **BAO-only analysis** (no CMB, no supernovae). *Impact: High.* DESI's full 3–4σ claim uses combined data. Our BAO-only analysis tests a weaker claim.

10. **Point estimates for supernova constraints** (in cross-dataset analysis). *Impact: Medium.* We use published best-fit (w_0, w_a) rather than full posteriors.

11. **Optimizer convergence**. *Impact: Low.* L-BFGS-B with reasonable bounds reliably finds the minimum for this smooth, low-dimensional problem.

13 Known Limitations

1. **Reduced statistical power from data-splitting.** By using only 6 of 13 measurements for testing, we lose power. Power calibration shows the median E_{split} under H_1 is only ≈ 2.7, confirming this limitation is severe.

2. **Data-split particularly underpowered for w_a.** The CPL parameterization means the low-z training set ($z < 1$) has only 23–48% leverage on the w_a parameter, which is where DESI's signal primarily resides.

3. **LOO average is conservative.** Averaging (rather than multiplying) LOO e-values sacrifices statistical power to maintain validity. The LOO average $E \approx 10$ is a lower bound on the evidence that could be extracted from a valid LOO procedure.

4. **No systematic error budget.** We treat the covariance matrix as exact. Unknown systematics could broaden error bars or introduce biases.

5. **Simplified cross-dataset analysis.** Using point estimates for supernova constraints (rather than full likelihoods with their own covariance) is approximate.

6. **No model-averaging or Bayesian comparison.** We compare two specific models. There may be other parametrizations that better capture any real dark energy evolution.

14 Comparison to Other Analyses

Analysis	Method	Finding	Conclusion
DESI DR2 (official)	Frequentist $\Delta\chi^2$	2.8σ ($k = 2$)	Dynamic DE
This work	AIC / BIC	$\Delta\mathrm{AIC} = 7.9$, $\Delta\mathrm{BIC} = 6.8$	Favors w_0w_aCDM
Ong et al. (2025)	Bayesian evidence	$\ln B = -0.57$	Favors ΛCDM
Wang & Mota (2025)	Tension metrics	2.95σ tension	Datasets inconsistent
This work	**E-values**	$E \approx 10$–15	**Moderate evidence**
This work	**Cross-dataset**	$E = 0.19$ **(DES-Y5)**	**Dataset tension**

The picture that emerges: moderate evidence for w_0w_aCDM within DESI BAO alone ($E \approx 10$–15), but this is undermined by cross-dataset tension. Three independent approaches (Bayesian model comparison, tension metrics, and e-values) agree that resolving inter-dataset tensions is the priority.

Part VII
Summary

15 What We Did

1. Loaded official DESI DR2 BAO data: 13 measurements of cosmic distances at 7 redshifts.
2. Computed theoretical distance predictions under ΛCDM ($w = -1$, no dark energy evolution) and w_0w_aCDM (dark energy evolves).
3. Used four e-value methods to assess evidence:
   - Maximized likelihood ratio (not a valid e-value, for reference only)
   - Uniform mixture e-value (valid, moderate evidence: $E \approx 15$)
   - Data-split e-value (valid but underpowered: $E = 1.4$)
   - LOO average e-value (valid, moderate evidence: $E \approx 10$)
4. Performed cross-dataset validation using supernova constraints.
5. Calibrated data-split power via Monte Carlo simulation.
6. Checked DR1→DR2 temporal stability.

16 What We Found

1. **Cross-dataset tension is the strongest finding.** DES-Y5's parameters predict DESI data *worse* than ΛCDM ($E = 0.19$), while Pantheon+'s predict it well ($E = 2049$). This $\sim$10,000× asymmetry indicates that the supernova catalogs disagree on what dark energy dynamics look like, raising the possibility that w_0w_aCDM is absorbing inter-dataset tension.

2. **Valid e-values converge on moderate evidence.** Two independent valid methods give consistent results: the uniform mixture e-value ($E \approx 15$) and the LOO average ($E \approx 10$) both indicate moderate evidence for $w_0 w_a$CDM over ΛCDM in DESI BAO data alone.

3. **The data-split e-value ($E = 1.4$) is inconclusive.** Power calibration shows that even when $w_0 w_a$CDM is the true model, the data-split test yields median $E \approx 2.7$. The observed $E = 1.4$ cannot distinguish between hypotheses. The data-split is underpowered for w_a because the low-z training set has only fractional leverage on the time-evolution parameter.

4. **The proper frequentist significance is 2.8σ.** The $\chi^2(2)$ distribution gives $p = 0.0026$ for $\Delta\chi^2 = 11.9$, not the 3.5σ from the incorrect 1-d.o.f. formula.

5. **Invalid statistics overstate the evidence.** The maximized likelihood ratio (LR = 392) and the LOO product ($E = 1062$) are not valid e-values. The former has $\mathbb{E}[E|H_0] = \infty$; the latter uses dependent folds. Both should be treated as descriptive statistics only.

17 What This Means

The evidence for dynamical dark energy from DESI BAO data is moderate ($E \approx 10$–15, corresponding roughly to 2.2–2.3σ) but undermined by cross-dataset inconsistency. This does *not* prove that dark energy is constant — it means:

- The evidence is neither absent nor compelling. Valid e-values consistently give $E \approx 10$–15, which is moderate evidence that should be taken seriously but does not constitute a discovery.
- The cross-dataset tension (DES-Y5 vs. Pantheon+ vs. DESI) is the more fundamental problem. Until this is resolved, it remains unclear whether $w_0 w_a$CDM is detecting real physics or absorbing systematic tensions.
- More data is needed. Key next steps include:
- DESI DR3+ (∼2027): more data, smaller errors
- Resolution of inter-dataset tensions (especially DESI vs. DES-Y5)
- Independent confirmation from Euclid and the Roman Space Telescope

---

Code and data: `https://github.com/jinyoungkim927/desi-evalue-analysis`
DESI data: `https://github.com/CobayaSampler/bao_data`

J. Kim, February 2026