Statistical Analysis of Animal Longevity Across Vertebrate Classes

Tim Wilcoxson

February 2026

Project 2 -- Data and Statistical Reasoning

Dataset: AnAge -- The Animal Ageing and Longevity Database

Source: https://genomics.senescence.info/species/

1. Overview

This report presents a statistical analysis of the AnAge database (de Magalhaes & Costa, 2009), a curated collection of longevity and life-history data for 4,645 species maintained by the Human Ageing Genomic Resources project (HAGR, 2023). The primary research question is whether maximum lifespan differs significantly across five major vertebrate classes: Mammalia, Aves, Teleostei, Reptilia, and Amphibia. A secondary analysis quantifies the allometric relationship between body size and longevity.

The analytical workflow follows the initial data analysis (IDA) framework for reproducible research (Lusa et al., 2024): exploratory screening of distributional properties and assumption violations precedes formal inference. Two inferential approaches -- a parametric baseline (one-way ANOVA) and a non-parametric alternative (Kruskal-Wallis H-test) -- are compared to demonstrate how assumption violations affect conclusions and to justify the final method selection. All code, data, and figures are available in the accompanying Jupyter notebook (analysis.ipynb).

2. Dataset Description

The AnAge database (version 2024, accessed February 2026) contains records for 4,645 species across 31 variables, including taxonomic classification (Kingdom through Species), life-history traits (maturity age, gestation period, litter size, birth weight, adult weight), and longevity metrics (maximum longevity in years, mortality rate parameters). The database was compiled and is maintained by de Magalhaes and Costa (2009) as part of the HAGR initiative at the University of Liverpool.

For this analysis, the dataset was filtered to five major vertebrate classes containing 4,396 species (95% of the database): Aves (1,513 species), Mammalia (1,349), Teleostei (806), Reptilia (547), and Amphibia (181). After removing records with missing longevity values, 3,909 species were available for hypothesis testing. The primary analysis variables are maximum longevity in years (numeric response variable) and taxonomic class (categorical grouping variable). Adult weight in grams was used as a secondary numeric predictor for the allometric analysis (n = 3,131 species with both measurements).

Data provenance: 85% of records carry an 'acceptable' quality rating from the HAGR curation team, 2% are rated 'high,' 11% 'low,' and 2% 'questionable.' Approximately 45% of specimens originate from wild populations and 41% from captive settings, with 14% of unknown origin. This mix of captive and wild data is an important caveat discussed in Section 7.

3. Experimental Design and Analytical Approach

The analytical workflow follows a structured, two-phase design motivated by the IDA framework

(Lusa et al., 2024), which advocates systematic screening of data properties before conducting planned statistical analyses. This ensures that the choice of inferential method is informed by actual data characteristics rather than default assumptions.

Phase 1: Baseline Exploration (Descriptive)

Descriptive statistics and visualizations establish a baseline understanding of the data. Summary statistics (mean, median, standard deviation) characterize central tendency and dispersion for each vertebrate class. Five visualizations -- histograms, box plots, scatter plots, bar charts, and Q-Q plots (Midway, 2020) -- reveal distributional shape, between-group differences, allometric relationships, sampling patterns, and assumption violations.

Phase 2: Inferential Testing (Baseline vs. Selected)

The inferential phase compares two approaches to the same hypothesis test. The baseline approach is the standard one-way ANOVA, which compares group means and assumes normally distributed data with equal variances across groups. The selected approach is the Kruskal-Wallis H-test (Kruskal & Wallis, 1952), a non-parametric alternative that compares rank distributions without distributional assumptions.

This comparison is meaningful for three reasons. First, the IDA screening phase revealed three critical assumption violations that undermine the parametric baseline: (1) severe non-normality (confirmed by Q-Q plots showing heavy right-tail departure in all classes; see Figure 5), (2) unequal variances (Levene's test $W = 12.45$, $p < 0.001$), and (3) highly unbalanced group sizes ranging from $n = 162$ (Amphibia) to $n = 1{,}394$ (Aves). Second, comparing both approaches demonstrates how assumption violations affect the magnitude of the detected effect: the parametric ANOVA underestimates the between-class signal because extreme outliers inflate within-group variance. Third, showing that both methods reach the same qualitative conclusion (reject H0) validates the robustness of the finding regardless of analytical method.

4. Methods

Descriptive Statistics

Central tendency (mean, median) and dispersion (standard deviation, interquartile range) were computed for maximum longevity within each vertebrate class. Both measures of central tendency are reported because the mean is sensitive to extreme values in skewed distributions, while the median provides a robust estimate of the typical value (Lusa et al., 2024). Frequency counts for categorical variables (data quality ratings, specimen origin) characterize the composition and provenance of the dataset.

Visualizations

Five visualizations were produced following principles for effective statistical graphics (Midway, 2020): (1) paired histograms compare raw and log10-transformed longevity distributions, demonstrating the necessity of transformation; (2) box plots compare longevity across classes, revealing medians, spreads, and outliers; (3) a scatter plot with OLS regression line quantifies the allometric weight-longevity relationship; (4) a horizontal bar chart of the top 10 taxonomic orders reveals sampling bias; and (5) Q-Q plots for all five classes assess normality, with Shapiro-Wilk test annotations quantifying the departure.

Hypothesis Test Selection and Justification

The null hypothesis states that the distribution of maximum longevity is identical across all five vertebrate classes (H0). The alternative states that at least one class differs (H1). Significance level: alpha = 0.05.

As a baseline, one-way ANOVA was applied. This is the standard parametric test for comparing means across k > 2 groups. However, ANOVA requires three assumptions (McDonald, 2014): (1) independence of observations, (2) normality within each group, and (3) homogeneity of variance (homoscedasticity). Assumptions (2) and (3) were tested and found to be violated:

- Non-normality: Q-Q plots for all five classes show systematic right-tail departure from the normal reference line (Figure 5), confirmed by Shapiro-Wilk tests (Shapiro & Wilk, 1965; W ranged from 0.64 for Teleostei to 0.88 for Aves, all p < 0.001). The raw longevity distribution has a mean of 25.5 years vs. a median of 15.0 years, indicating severe positive skew.
- Unequal variances: Levene's test (Levene, 1960) rejected the null hypothesis of equal variances (W = 12.45, p = 4.59 x 10^-10), confirming heteroscedasticity. Standard deviations range from 10.7 (Amphibia) to 21.6 (Teleostei).
- Unbalanced groups: Sample sizes span a nearly 9:1 ratio (162 to 1,394). Combined with heteroscedasticity, this imbalance makes the ANOVA F-test unreliable.

Given these violations, the Kruskal-Wallis H-test (Kruskal & Wallis, 1952) was selected as the primary inferential method. This non-parametric test compares rank distributions rather than group means, making it robust to non-normality, unequal variances, and outliers. The trade-off is reduced statistical power compared to ANOVA when parametric assumptions hold -- but in this dataset, those assumptions do not hold.

Post-hoc Comparisons

When the omnibus test rejects H0, pairwise Mann-Whitney U tests (Mann & Whitney, 1947) identify which specific class pairs differ. With k = 5 groups there are C(5,2) = 10 comparisons. The Bonferroni correction (Dunn, 1961) adjusts the significance threshold to alpha/10 = 0.005, controlling the family-wise Type I error rate. This correction is conservative: it minimizes false positives at the cost of increased Type II error (missed true differences).

Effect Sizes

Two effect size measures are reported to enable direct comparison between approaches. For

ANOVA: eta-squared (eta-sq = SS_between / SS_total), the proportion of total variance explained by group membership. For Kruskal-Wallis: epsilon-squared (epsilon-sq = H / (n - 1)), the non-parametric analogue estimating the proportion of variance in ranks explained by group membership. Conventional benchmarks (Cohen, 1988; Tomczak & Tomczak, 2014): ~0.01 = small, ~0.06 = medium, ~0.14 = large.

5. Results and Performance Evaluation

Descriptive Findings

Across all 4,141 species with longevity data (3,909 in the five target vertebrate classes), the overall mean maximum longevity was 25.5 years with a median of 15.0 years, confirming severe right-skewness. Among the five vertebrate classes, Reptilia had the highest median longevity (17.8 years), followed by Mammalia (17.0), Aves (14.6), Amphibia (11.9), and Teleostei (10.0). However, Teleostei exhibited the widest variability (SD = 21.6 years), reflecting diversity from short-lived tropical fish to sturgeon exceeding 200 years.

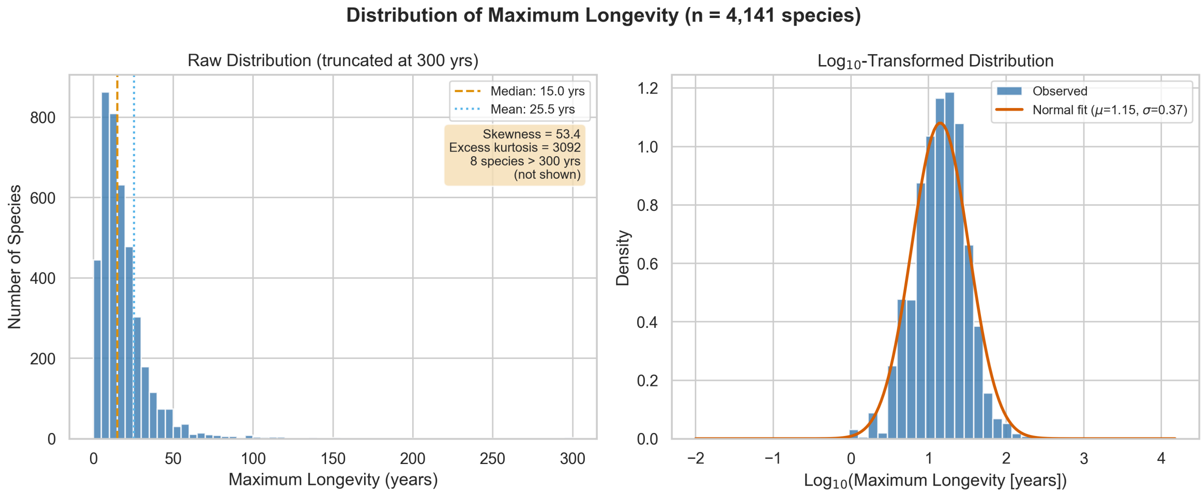

Figure 1. Distribution of maximum longevity: raw (left) and log10-transformed (right). The raw histogram shows extreme right-skew; the log transformation reveals an approximately bell-shaped pattern.

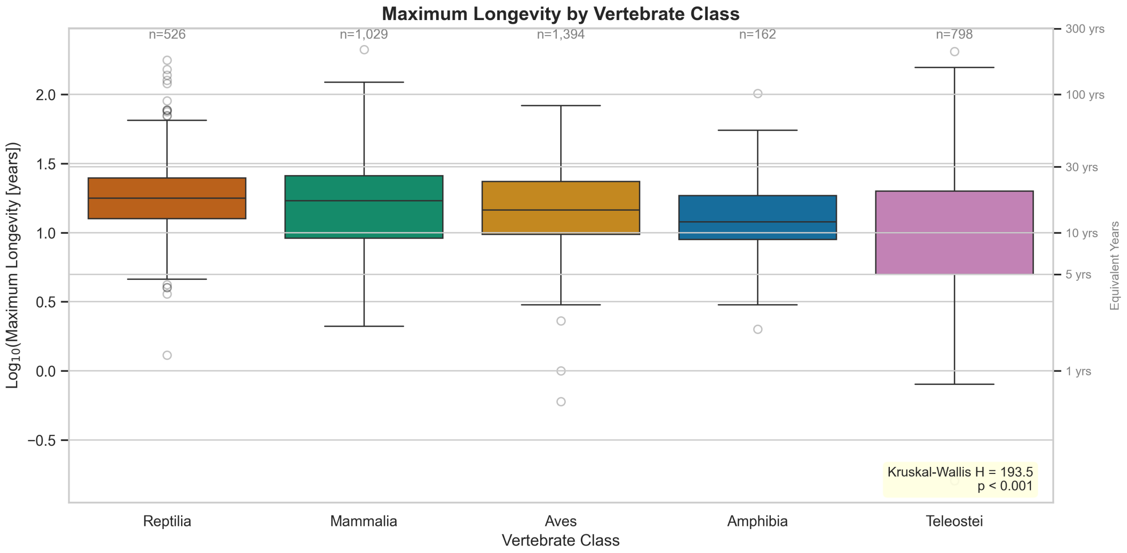

Figure 2. Box plots of log10(maximum longevity) by vertebrate class, ordered by median. Sample sizes annotated above each box.

Allometric Relationship

The scatter plot of log10(adult weight) vs. log10(maximum longevity) across 3,131 vertebrate species yielded a Pearson correlation coefficient of r = 0.5678 (p < 0.001), indicating a moderate positive linear association on the log-log scale (Figure 3). This confirms the well-established allometric principle (Speakman, 2005) that larger-bodied species tend to live longer, though the relationship explains approximately 32% of the variance (r-squared = 0.3224), leaving substantial residual variation attributable to factors such as metabolic rate, predation pressure, and reproductive strategy.

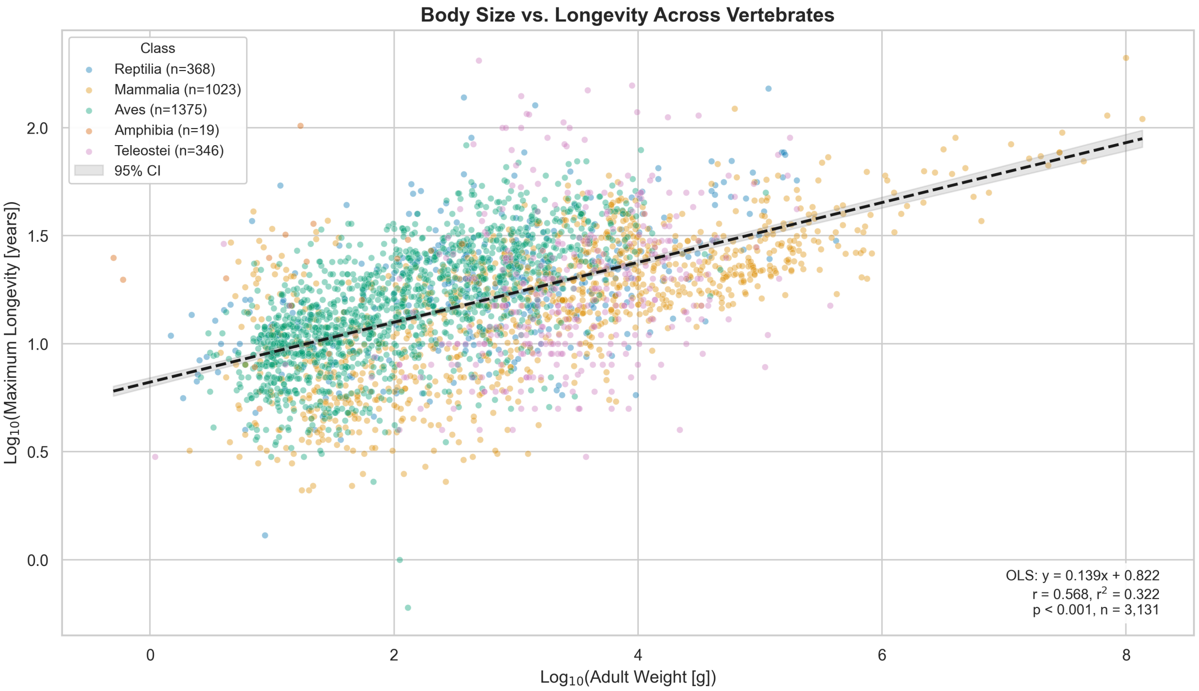

Figure 3. Allometric scaling of body size and longevity across vertebrate classes (n = 3,131). Dashed line: OLS regression (r = 0.5678).

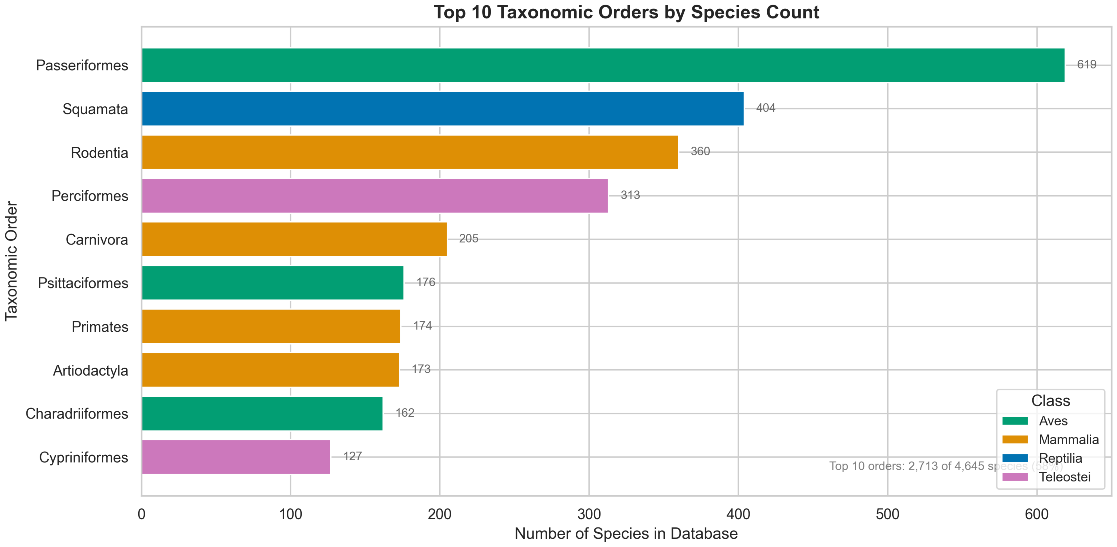

Figure 4. Top 10 taxonomic orders by species count in the AnAge database, colored by class. Reveals sampling bias toward well-studied taxa such as Passeriformes (619 species).

Baseline Results: One-Way ANOVA (Parametric)

The one-way ANOVA yielded F(4, 3904) = 12.67 with p = 3.03 x 10^-10, rejecting H0 at alpha = 0.05. The effect size was eta-squared = 0.0128 (small), indicating that class membership explains approximately 1.3% of the total variance in raw longevity values. However, this result must be interpreted with caution because both the normality and equal-variance assumptions are violated (see Section 4). The low effect size is partly an artifact of the extreme right-skew: a small number of exceptionally long-lived species inflate within-group variance, which in turn suppresses the between-group signal in the ANOVA F-ratio.

Selected Results: Kruskal-Wallis H-Test (Non-Parametric)

The Kruskal-Wallis H-test yielded H(4) = 193.51 with p = 9.33 x 10^-41 (n = 3,909 species across 5 classes), decisively rejecting H0. The effect size was epsilon-squared = 0.0495 (small), indicating that class membership explains approximately 5% of the variance in ranked longevity. This is the primary and most reliable result of the hypothesis test, as the Kruskal-Wallis test does not depend on the violated assumptions.

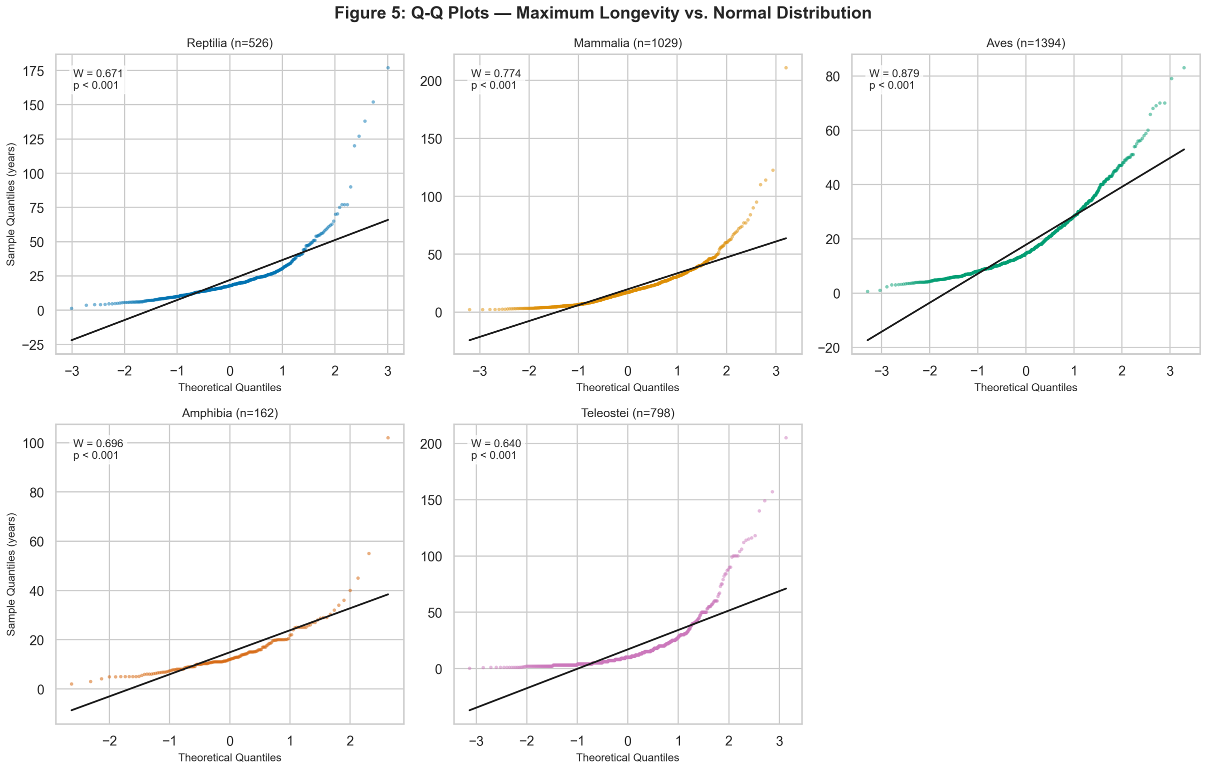

Figure 5. Q-Q plots for all five vertebrate classes with Shapiro-Wilk test statistics (all W < 0.88, p < 0.001). Systematic right-tail departure from the normal line confirms non-normality.

Performance Comparison: Parametric vs. Non-Parametric

Comparing the two approaches reveals important methodological insights. Both tests agree on the qualitative conclusion (reject H0; p < 0.001), confirming that the finding of between-class differences is robust to method selection. However, they differ in the magnitude of the detected effect:

- ANOVA eta-squared = 0.0128 vs. Kruskal-Wallis epsilon-squared = 0.0495. Although these measures are not directly comparable (eta-squared uses raw values; epsilon-squared uses ranks), the non-parametric approach captures approximately 4 times more between-class signal because rank-based analysis neutralizes the outlier-driven variance inflation that suppresses the ANOVA effect size.

- The discrepancy arises because ANOVA operates on raw values, where extreme outliers (species with recorded maximum longevities exceeding 200 years) inflate within-group variance and suppress the F-ratio. The Kruskal-Wallis test operates on ranks, neutralizing outlier influence and revealing the between-class pattern more clearly.
- ANOVA p = 3.03 x 10^-10 vs. Kruskal-Wallis p = 9.33 x 10^-41. The non-parametric test achieves a p-value 31 orders of magnitude smaller, reflecting its greater sensitivity to rank-order differences in the presence of skew.

This comparison demonstrates why method selection matters. A naive application of ANOVA to this dataset would produce a valid rejection of H0 but substantially underestimate the strength of the between-class signal. The IDA framework (Lusa et al., 2024) prevented this by ensuring that assumption checking preceded method selection.

Post-hoc Pairwise Comparisons

After Bonferroni correction (adjusted alpha = 0.005), 9 of 10 class pairs showed statistically significant differences in longevity. The sole non-significant pair was Aves vs. Mammalia (adjusted p = 0.275), suggesting that birds and mammals share similar longevity distributions despite fundamental differences in physiology and body plan. The most divergent pair was Reptilia vs. Teleostei (adjusted p = 1.92 x 10^-30), consistent with the large median gap between these classes (17.8 vs. 10.0 years). Rank-biserial correlations (effect sizes for each pair) ranged from |r_rb| = 0.05 (Aves vs. Mammalia) to |r_rb| = 0.38 (Reptilia vs. Teleostei), confirming that even statistically significant pairs often exhibit modest practical effect sizes.

A methodological trade-off: the Bonferroni correction is the most conservative multiple-comparison method, strictly controlling the family-wise Type I error rate. It is possible that the one non-significant result (Aves vs. Mammalia) represents a genuine small difference masked by overcorrection. Less conservative alternatives such as the Holm step-down procedure (Holm, 1979) or the Benjamini-Hochberg false discovery rate control (Benjamini & Hochberg, 1995) could be explored in future work.

Confidence Intervals

Bootstrap 95% confidence intervals (Efron, 1979; percentile method, 9,999 resamples) quantify uncertainty in the median longevity estimates. Reptilia: 17.8 years [17.0, 19.0]; Mammalia: 17.0 [16.2, 17.9]; Aves: 14.6 [13.8, 15.0]; Amphibia: 11.9 [11.0, 13.6]; Teleostei: 10.0 [9.0, 11.0]. The non-overlapping CIs for Reptilia/Mammalia vs. Teleostei corroborate the hypothesis test findings.

For the allometric correlation: Pearson r = 0.5678, 95% CI [0.5436, 0.5911] (Fisher z-transformation, n = 3,131). The narrow interval confirms this is a precisely estimated, moderate positive association. For the Kruskal-Wallis effect size: epsilon-squared = 0.0495, 95% CI [0.0366, 0.0662] (bootstrap, 5,000 resamples), confirming the small-effect classification (Cohen, 1988).

Sensitivity Analysis

To assess robustness, the Kruskal-Wallis test was re-run on four data subsets: (1) the full dataset (H = 193.51, p = 9.33e-41, epsilon-sq = 0.0495); (2) excluding low-quality records (H = 194.10, p = 6.97e-41, epsilon-sq = 0.0510); (3) wild specimens only (H = 97.06, p = 4.15e-20, epsilon-sq = 0.0512); and (4) captive specimens only (H = 132.51, p = 1.13e-27, epsilon-sq = 0.0697). All four subsets reject H0 with similar effect sizes, confirming that the between-class longevity difference is not an artifact of data quality ratings or specimen origin.

6. Interpretation for Non-Technical Audience

Different types of animals really do live different lengths of time, and this is not just coincidence -- it is a real biological pattern. We compared the maximum lifespans of nearly 4,000 species across five major groups of vertebrates (mammals, birds, bony fish such as salmon and goldfish, reptiles, and amphibians) and found strong evidence that the group an animal belongs to is meaningfully connected to how long it can live.

Reptiles tend to have the longest lifespans among these groups, with a typical maximum around 18 years, while bony fish have the shortest at about 10 years. Interestingly, birds and mammals turned out to be very similar in their lifespan patterns -- around 15 to 17 years for a typical species -- even though they are very different kinds of animals.

We also confirmed that bigger animals generally live longer. Body size explains roughly a third of the variation in how long species live -- on a scale from 0 (no relationship) to 1 (perfect relationship), the correlation scored 0.57, indicating a moderately strong connection. The remaining two-thirds of the variation depends on other factors like diet, habitat, predation risk, and reproductive strategy.

We tested this question using two different statistical methods -- a standard approach and a more robust alternative -- and both gave the same answer: the differences are real (the probability of seeing these results by chance alone is far less than 1 in a million). The robust method was actually better at detecting the differences because it handles unusual data more effectively. However, while the differences between groups are statistically real, the group label alone explains only about 5% of the variation in lifespan. Most of the variation comes from differences among species within each group.

7. Limitations, Potential Bias, and Ethical Considerations

Dataset Limitations

- Captive vs. wild specimen bias: Approximately 41% of longevity records come from captive specimens. Captive animals are protected from predation, disease, and starvation, which may inflate maximum longevity estimates relative to wild populations. This is particularly pronounced for large mammals and birds commonly kept in zoos.

- Missing data: 19 of 31 variables contain missing values. Metabolic rate is only 14% complete, preventing metabolic scaling analysis. Even the primary response variable (maximum longevity) is missing for 10.9% of species. These missing records are unlikely to be random -- species with shorter lifespans in remote habitats may be systematically understudied (Lusa et al., 2024).

Measurement Bias

- Maximum longevity is an extreme-value statistic that is systematically harder to capture for species with small populations, short generation times, or elusive behavior. Teleostei and Amphibia are particularly affected: many fish species are monitored only through catch records, and amphibians in tropical regions may never be recaptured after initial marking. As a result, the recorded maximum longevity for these taxa likely underestimates their true biological potential, compressing between-class differences.
- Phylogenetic non-independence: Species sharing recent common ancestry are not statistically independent observations (Felsenstein, 1985). The Kruskal-Wallis test assumes independent samples, but closely related species within each class will have correlated longevity values. This inflates the effective sample size and may produce artificially narrow confidence intervals. Phylogenetic comparative methods (e.g., phylogenetic generalized least squares) would address this limitation but require a resolved phylogeny for all 3,909 species.

Sampling and Representation Bias

- Taxonomic overrepresentation: The database heavily favors well-studied taxa. Passeriformes (songbirds) alone contributes 619 species, while entire classes like Amphibia have only 181 entries. Results are therefore more statistically reliable for Aves and Mammalia than for Amphibia, where the smaller sample reduces power.
- Survivorship bias: Only species with published longevity records are included. Recently discovered, rare, or data-deficient species are systematically excluded, potentially skewing toward well-documented taxa that may not represent the full diversity of vertebrate lifespans.
- Geographic bias: Species from well-funded research regions (North America, Europe, Australasia) are overrepresented compared to tropical and developing-world fauna, limiting global generalizability.

Ethical Considerations

- Conservation policy risk: If longevity data were used to inform conservation resource allocation, the taxonomic and geographic biases could lead to misallocation of resources -- prioritizing already well-studied species while neglecting data-deficient but ecologically critical taxa. Decision-makers should account for data completeness before drawing policy conclusions.
- Data quality transparency: While 85% of records are rated 'acceptable' by the HAGR curation

team, 13% are rated 'low' or 'questionable.' Presenting statistical results without acknowledging these quality tiers could overstate precision and reliability.

- Captive-specimen ethics: The use of captive-derived longevity data implicitly depends on animals maintained in zoos and research facilities. Any downstream application of these findings should consider the ethical dimensions of captive animal data sourcing.

8. Conclusion

This analysis examined maximum longevity across 3,909 vertebrate species from the AnAge database, comparing five major classes. Both parametric (ANOVA) and non-parametric (Kruskal-Wallis) approaches confirmed statistically significant between-class differences in longevity (H(4) = 193.51, p = 9.33 x 10^-41), with 9 of 10 class pairs differing significantly after Bonferroni correction. The sole exception was Aves vs. Mammalia, which share similar longevity distributions.

The small effect size (epsilon-squared = 0.0495) indicates that class membership explains approximately 5% of ranked longevity variance, with within-class variation dominating. The allometric relationship between body size and longevity (r = 0.5678, n = 3,131) confirms established scaling theory. Sensitivity analyses across four data subsets demonstrated that these findings are robust to data quality and specimen origin.

Key limitations include captive-specimen bias, taxonomic overrepresentation of mammals and birds, and phylogenetic non-independence among related species. Future work should incorporate phylogenetic comparative methods to account for shared evolutionary history.

Software Environment

All analyses were conducted in Python 3.13 using SciPy 1.17.0, Pandas 3.0.1, NumPy 2.4.2, Matplotlib 3.10.8, and Seaborn 0.13.2. The PDF report was generated with fpdf2 2.8.6. The complete analysis notebook and all data files are available in the project repository.

9. References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates.

de Magalhaes, J. P., & Costa, J. (2009). A database of vertebrate longevity records and their relation to other life-history traits. Journal of Evolutionary Biology, 22(8), 1770-1774. https://doi.org/10.1111/j.1420-9101.2009.01783.x

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56(293), 52-64. https://doi.org/10.1080/01621459.1961.10482090

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), 1-26. https://doi.org/10.1214/aos/1176344552

Felsenstein, J. (1985). Phylogenies and the comparative method. The American Naturalist, 125(1), 1-15. https://doi.org/10.1086/284325

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65-70.

Human Ageing Genomic Resources [HAGR]. (2023). AnAge: The Animal Ageing and Longevity Database. University of Liverpool. https://genomics.senescence.info/species/

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583-621. https://doi.org/10.2307/2280779

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), Contributions to Probability and Statistics (pp. 278-292). Stanford University Press.

Lusa, L., Proust-Lima, C., Schmidt, C. O., Lee, K. J., le Cessie, S., Baillie, M., Lawrence, F., & Huebner, M. (2024). Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis. PLoS ONE, 19(5), e0295726. https://doi.org/10.1371/journal.pone.0295726

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 18(1), 50-60. https://doi.org/10.1214/aoms/1177730491

McDonald, J. H. (2014). Handbook of Biological Statistics (3rd ed.). Sparky House Publishing.

Midway, S. R. (2020). Principles of effective data visualization. Patterns, 1(9), 100141. https://doi.org/10.1016/j.patter.2020.100141

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3-4), 591-611. https://doi.org/10.1093/biomet/52.3-4.591

Speakman, J. R. (2005). Body size, energy metabolism and lifespan. Journal of Experimental Biology, 208(9), 1717-1730. https://doi.org/10.1242/jeb.01556

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Trends in Sport Sciences, 1(21), 19-25.