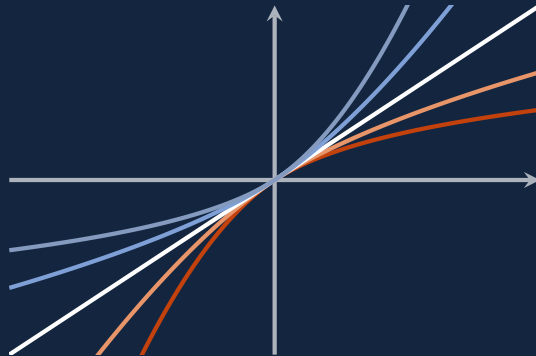




UNIVERSITÀ DI PISA

STATISTICS FOR DATA SCIENCE

power transformations • likelihood • inference



$$\psi(x, \lambda) = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda} & x \geq 0 \\ -\frac{(1-x)^{2-\lambda} - 1}{2-\lambda} & x < 0 \end{cases}$$

Yeo–Johnson Power Transformations: From Theory to Statistical Inference

Project Report

Giuseppe Muschetta

M.Sc. in Data Science & Business Informatics

Course Instructor: Prof. Salvatore Ruggieri

Academic Year: 2024–2025

Abstract

Many classical procedures, from the t -test to linear regression, rely on approximately normal data, yet real measurements are often skewed. Power transformations address the problem, but the classical Box–Cox family is defined only for strictly positive values. This project studies the family introduced by Yeo and Johnson [11], which is defined on the whole real line, and implements it from scratch in R together with the tools needed to use it for inference: the log-likelihood with its Jacobian term, maximum-likelihood estimation with the L-BFGS-B optimizer, an asymptotic covariance matrix from a numerical Hessian, and a likelihood-ratio test. The work proceeds in three steps. First, a strongly skewed mixture shows that the earlier signed power transformation can leave a bimodal distribution even at its best parameter. Second, the Yeo–Johnson transformation is applied to the same mixture and produces a single-peaked, nearly normal shape. Third, the full inferential analysis is carried out on Darwin’s classic data on cross- and self-fertilised plants: the estimate $\hat{\lambda} = 1.305$, the parameter estimates, the covariance matrix and the 0.01-quantile of the mean difference all reproduce the values published in the original paper.

Contents

1	Introduction	2
2	Power transformation families	2
2.1	The Box–Cox transformation	2
2.2	The signed power transformation	2
2.3	From the modified modulus to the Yeo–Johnson family	3
3	A skewed mixture: why a new family is needed	4
3.1	The test distribution	4
3.2	Choosing the parameter	4
3.3	Signed power: a bimodal result	4
3.4	Yeo–Johnson: a single peak	5
4	Likelihood-based estimation and inference	5
4.1	The model	5
4.2	The log-likelihood	6
4.3	Estimation	6
4.4	Testing whether a transformation is needed	6
4.5	Back to the original scale	6
5	Application: Darwin’s plant experiment	6
5.1	The data	6
5.2	Analysis on the original scale	7
5.3	Maximum-likelihood estimates	7
5.4	Effect of the transformation	7
5.5	Is the transformation necessary?	7
5.6	Conclusion on the original scale	7
5.7	Comparison with the published results	8
6	Implementation	8
7	Conclusions	9
A	Core functions	11

1 Introduction

The normality assumption is behind a large share of classical statistics: the paired and two-sample t -tests, the analysis of variance and the inference attached to linear regression all assume, at least approximately, normal errors. When data are skewed these procedures can lose validity or power, and a natural remedy is to transform the data so that, on the new scale, the assumption becomes plausible.

Power transformations are the standard tool for this purpose. The family proposed by Box and Cox [2] is the best known, but it is defined only for strictly positive observations. Data that take both signs, such as paired differences, changes over time or residuals, need a shift before the transformation can be applied, and the choice of the shift is arbitrary. Yeo and Johnson [11] proposed a family that is defined on the whole real line, coincides with a Box–Cox transformation on the shifted positive values, and treats negative values symmetrically.

This report documents a project that studies the Yeo–Johnson family from its definition to its use in inference. All transformations, likelihood functions and diagnostics were written in R [8] without relying on packages that already implement them; only `ggplot2` [10] and `gridExtra` are used for graphics. The project is organised in three parts, which correspond to the three main scripts:

Part 1. The limits of earlier transformations (`1st_part.R`): the signed power transformation of Bickel and Doksum [1] is applied to a strongly skewed mixture distribution.

Part 2. The Yeo–Johnson family (`2nd_part.R`): its construction, its geometry and its behaviour on the same mixture.

Part 3. Inference on Darwin’s data (`3rd_part.R`): maximum-likelihood estimation, normality diagnostics, a likelihood-ratio test and a quantile estimate on the experiment analysed in the original paper.

A fourth file, `functions.R`, collects the shared functions. Section 2 introduces the transformation families, Section 3 reports the experiments on the mixture, Section 4 derives the likelihood and the estimation procedure, Section 5 presents the analysis of Darwin’s data and compares it with the published results, Section 6 describes the implementation and Section 7 concludes.

2 Power transformation families

2.1 The Box–Cox transformation

For a strictly positive observation x , Box and Cox [2] defined

$$\psi_{BC}(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log x, & \lambda = 0. \end{cases} \quad (1)$$

The subtraction of 1 and the division by λ make the family continuous in λ : as $\lambda \rightarrow 0$, $(x^\lambda - 1)/\lambda \rightarrow \log x$. Values $\lambda < 1$ give concave transformations, which compress large values and reduce right skewness; values $\lambda > 1$ give convex transformations, which reduce left skewness; $\lambda = 1$ is a simple shift and leaves the shape of the distribution unchanged. The restriction $x > 0$ is the main practical limitation of the family.

2.2 The signed power transformation

To handle observations of any sign, Bickel and Doksum [1] proposed the signed power transformation

$$\psi_{SP}(x, \lambda) = \frac{\text{sgn}(x) |x|^\lambda - 1}{\lambda}, \quad \lambda > 0. \quad (2)$$

It is defined on the whole real line, but it applies the same power to both signs. When the distribution is strongly skewed, the positive and negative parts of the data would need different corrections, and a single power can push them apart instead of bringing them together. Section 3.3 shows this effect on a concrete example.

2.3 From the modified modulus to the Yeo–Johnson family

Yeo and Johnson start from a modified modulus transformation with two parameters, one for each sign:

$$\psi(x; \lambda_+, \lambda_-) = \begin{cases} \frac{(x+1)^{\lambda_+} - 1}{\lambda_+}, & x \geq 0, \\ -\frac{(-x+1)^{\lambda_-} - 1}{\lambda_-}, & x < 0, \end{cases} \quad (3)$$

with the usual logarithmic limits when a parameter equals zero. The shift by one makes both branches vanish at $x = 0$ and keeps the transformation finite near the origin. Requiring the second derivative with respect to x to be continuous at $x = 0$, where the two branches meet, gives

$$(\lambda_+ - 1) = -(\lambda_- - 1) \iff \lambda_+ + \lambda_- = 2, \quad (4)$$

so a single parameter $\lambda = \lambda_+$ is enough and $\lambda_- = 2 - \lambda$.

Definition 1 (Yeo–Johnson transformation). For $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$,

$$\psi(x, \lambda) = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda}, & x \geq 0, \lambda \neq 0, \\ \log(x+1), & x \geq 0, \lambda = 0, \\ -\frac{(-x+1)^{2-\lambda} - 1}{2-\lambda}, & x < 0, \lambda \neq 2, \\ -\log(-x+1), & x < 0, \lambda = 2. \end{cases} \quad (5)$$

For non-negative values the transformation is the Box–Cox transformation of $x + 1$; for negative values it is the Box–Cox transformation of $-x + 1$ with parameter $2 - \lambda$, reflected through the origin. The transformation is invertible: for $\psi \geq 0$, $x = (\lambda\psi + 1)^{1/\lambda} - 1$ (or $e^\psi - 1$ when $\lambda = 0$), and for $\psi < 0$, $x = 1 - (1 - (2 - \lambda)\psi)^{1/(2-\lambda)}$ (or $1 - e^{-\psi}$ when $\lambda = 2$).

Proposition 1 (Yeo and Johnson, 2000). *The transformation (5) has the following properties.*

- (i) $\psi(x, \lambda) \geq 0$ for $x \geq 0$ and $\psi(x, \lambda) < 0$ for $x < 0$; moreover $\psi(x, 1) = x$.
- (ii) ψ is continuous in (x, λ) , including the special values $\lambda = 0$ and $\lambda = 2$.
- (iii) ψ is increasing in x for every λ , and increasing in λ for every $x \neq 0$.
- (iv) ψ is concave in x for $\lambda < 1$ and convex in x for $\lambda > 1$.
- (v) ψ is convex in λ for $x > 0$ and concave in λ for $x < 0$.

Property (iv) is what makes the family useful in practice: values $\lambda < 1$ correct right skewness and values $\lambda > 1$ correct left skewness, on the whole real line. Figure 1 reproduces the comparison of Figure 2 of the original paper. The scripts also check properties (ii) and (iii) numerically: `test_continuity` evaluates the transformation on a fine grid around a point and compares it with the value at the point, and `test_increasing` compares values after small increments of x and λ . Continuity held at $x = 0$ for $\lambda \in \{-1, 0, 0.5, 1, 2, 3\}$ and at the special values $\lambda = 0$ and $\lambda = 2$ for $x = \pm 1.5$; monotonicity in x and in λ held at all 30 tested pairs, $x \in \{-3, -1, -0.2, 0.2, 1, 3\}$ and $\lambda \in \{0, 0.5, 1, 1.5, 2\}$.

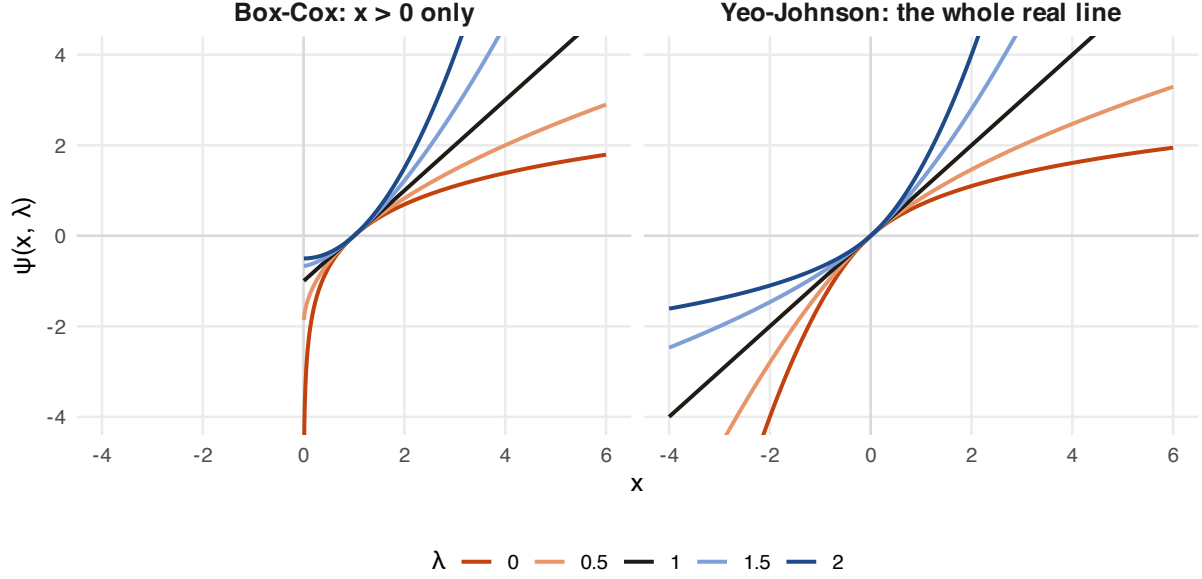


Figure 1: Box–Cox and Yeo–Johnson transformations for $\lambda \in \{0, 0.5, 1, 1.5, 2\}$. Box–Cox is defined only for $x > 0$; Yeo–Johnson extends smoothly to negative values, with concave curves for $\lambda < 1$ and convex curves for $\lambda > 1$ on the whole line, and the identity at $\lambda = 1$.

3 A skewed mixture: why a new family is needed

3.1 The test distribution

To compare the two families on a difficult case, the scripts use the mixture

$$f(x) = 0.3 \phi(x) + 0.7 g(x), \quad g(x) = \frac{1}{6}(x+2)^3 e^{-(x+2)} \mathbf{1}\{x > -2\}, \quad (6)$$

where ϕ is the standard normal density and g is a Gamma density with shape 4 and rate 1, shifted by -2 . The mixture takes both signs and is strongly asymmetric. A sample of 10,000 observations is drawn with replacement from a grid of 3,000 equally spaced points, with probabilities proportional to f ; the random seed is fixed (`set.seed(123)`) so that every number below can be reproduced.

3.2 Choosing the parameter

For each candidate λ , the sample is transformed and compared with a normal distribution that has the mean and standard deviation of the transformed values. The parameter is chosen by minimising a criterion based on the Kullback–Leibler divergence [7] between the transformed data and this normal model, with the bounded quasi-Newton method L-BFGS-B [3] on the interval $\lambda \in [0.01, 2]$. For the Yeo–Johnson transformation the transformed density is estimated with a Gaussian kernel. In Yeo and Johnson’s paper the same idea appears through the Kullback–Leibler information number, which they minimise to define the transformation that brings a distribution closest to normality.

3.3 Signed power: a bimodal result

For the signed power transformation the optimiser returns $\lambda^* = 0.237$, with transformed mean -1.588 and standard deviation 3.693 . Figure 2 (left) shows the outcome: even at its best parameter, the transformed distribution has two separate peaks. A single power applied to both signs pulls the negative observations away from the positive ones, so the transformation creates

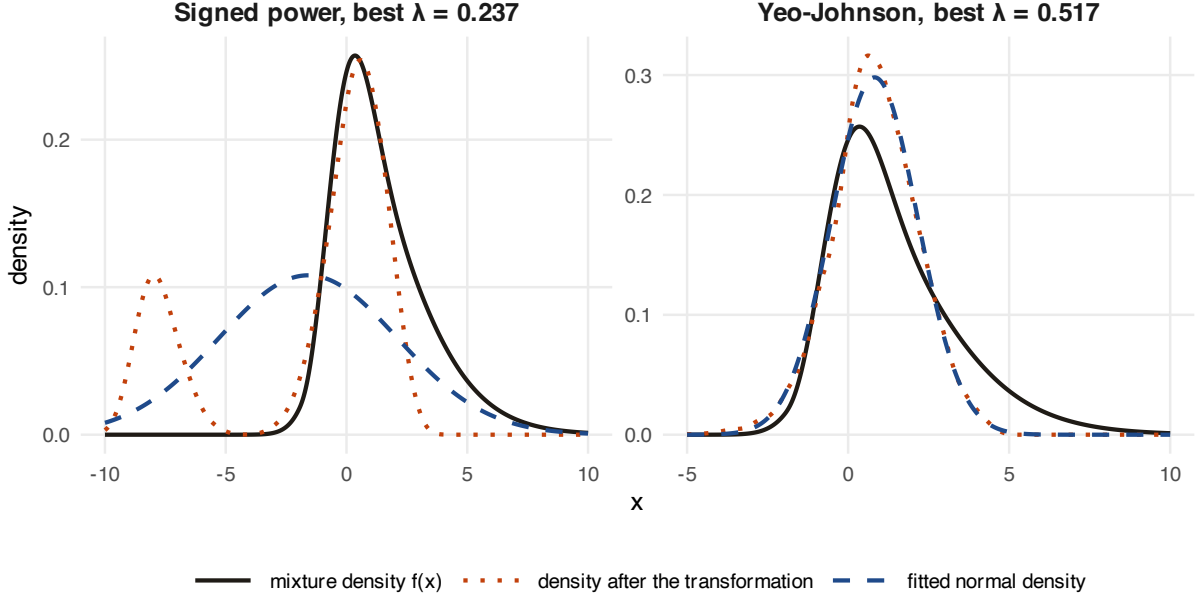


Figure 2: The mixture density f (solid), the density of the transformed sample at the best parameter (dotted) and the normal density with the same mean and standard deviation (dashed). Left: signed power transformation, which leaves two peaks. Right: Yeo–Johnson transformation, which gives a single, nearly normal peak.

a bimodality that the fitted normal curve cannot describe. This is the limitation that motivates a family with different behaviour on the two sides of zero.

3.4 Yeo–Johnson: a single peak

Applied to the same mixture, the Yeo–Johnson transformation reaches its best value at $\lambda^* = 0.517$, with transformed mean 0.823 and standard deviation 1.339. Figure 2 (right) shows a single-peaked distribution close to its fitted normal density, as in Figure 3 of the original paper. Because the negative branch uses the parameter $2 - \lambda$, a concave correction of the long right tail comes with a matching convex correction on the left, and the two halves of the distribution stay together.

4 Likelihood-based estimation and inference

4.1 The model

Following Yeo and Johnson, the transformation parameter is treated as an unknown parameter of a statistical model: for some λ , the transformed observations are independent and normally distributed,

$$\psi(X_i, \lambda) \sim N(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (7)$$

The density of the original observations follows from the change of variables, and the Jacobian is

$$\frac{\partial \psi(x, \lambda)}{\partial x} = (|x| + 1)^{\text{sgn}(x)(\lambda-1)}. \quad (8)$$

4.2 The log-likelihood

With $\boldsymbol{\theta} = (\lambda, \mu, \sigma^2)$, the log-likelihood of a sample $x_1, \dots, x_n$ is

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (\psi(x_i, \lambda) - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \text{sgn}(x_i) \log(|x_i| + 1). \quad (9)$$

The last term is the logarithm of the Jacobian. Without it, the likelihood would compare transformed data measured on different scales for different values of λ , and the estimate of λ would be meaningless. The function `log_likelihood` implements (9) and returns its negative, so that it can be minimised.

4.3 Estimation

For fixed λ , the maximum of (9) over μ and σ^2 is reached at the mean and the variance of the transformed data. The script uses this to proceed in two stages:

1. a one-dimensional search for λ with L-BFGS-B on $[-2, 2]$, starting from $\lambda = 0$, where μ and σ^2 are replaced by the sample mean and variance of $\psi(x_i, \lambda)$;
2. a joint optimisation of (λ, μ, σ^2) with L-BFGS-B, started from the first-stage values, which also returns the numerical Hessian $\hat{H}$ of $-\ell$ at the optimum.

The inverse of the Hessian estimates the covariance of $\hat{\boldsymbol{\theta}}$; the script reports $n \hat{H}^{-1}$, the estimated asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, which is the quantity given in the original paper.

4.4 Testing whether a transformation is needed

Since $\psi(x, 1) = x$, the hypothesis $H_0: \lambda = 1$ states that the data are already normal on the original scale. It is tested with the likelihood-ratio statistic

$$\Lambda = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(1, \tilde{\mu}, \tilde{\sigma}^2)] \sim \chi_1^2 \quad \text{under } H_0, \quad (10)$$

where $\tilde{\mu}$ and $\tilde{\sigma}^2$ are the estimates under $\lambda = 1$.

4.5 Back to the original scale

Inference made on the transformed scale can be carried back to the original one through the inverse transformation. Following Carroll and Ruppert [4], as in the original paper, the γ -th quantile of the mean difference is approximated by

$$\hat{q}_\gamma = \psi^{-1} \left(\hat{\mu} + t_{\gamma, n-1} \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\lambda} \right), \quad (11)$$

where $t_{\gamma, n-1}$ is the γ -th quantile of Student's t distribution with $n - 1$ degrees of freedom.

5 Application: Darwin's plant experiment

5.1 The data

Darwin [5] studied the effect of cross- and self-fertilisation on the growth of plants. Fifteen pairs of seedlings of the same age, one cross-fertilised and one self-fertilised, were grown together so that each pair shared nearly identical conditions; the experiment later became a classic example in Fisher's *The Design of Experiments* [6]. The data are the fifteen differences between the final heights of the two plants in each pair (Table 1). The question is whether cross-fertilised plants grow taller, that is, whether the mean difference is positive.

Pair	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Difference	-8.4	-6.0	0.7	1.0	1.8	2.0	2.9	3.0	3.5	3.6	5.1	6.1	7.0	7.5	9.3

Table 1: Differences in final height between the cross- and the self-fertilised plant of each pair, sorted.

5.2 Analysis on the original scale

A one-sided paired t -test of $H_0: \mu = 0$ against $H_1: \mu > 0$ gives $t = 2.142$ with 14 degrees of freedom and $p = 0.025$: at the 5% level the data support Darwin’s claim. The validity of the test, however, depends on the normality of the differences, and two observations, -8.4 and -6.0 , lie far below the others. The sample skewness statistic is -0.816 , and the Shapiro–Wilk test [9] gives $W = 0.899$ with $p = 0.093$. The left panels of Figure 3 show the skewness: a long left tail in the density and points falling away from the line at the lower end of the normal Q–Q plot.

5.3 Maximum-likelihood estimates

The procedure of Section 4 gives

$$\hat{\lambda} = 1.305, \quad \hat{\mu} = 4.570, \quad \hat{\sigma}^2 = 29.785, \quad (12)$$

with estimated asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$

$$n\hat{H}^{-1} = \begin{pmatrix} 0.434 & 3.270 & 26.278 \\ 3.270 & 54.445 & 198.175 \\ 26.278 & 198.175 & 3366.983 \end{pmatrix}. \quad (13)$$

The estimate $\hat{\lambda} > 1$ calls for a convex transformation, which is what left-skewed data need. The positive covariance between $\hat{\lambda}$ and $\hat{\mu}$ reflects the fact that a larger λ stretches the positive differences and moves the transformed mean upwards.

5.4 Effect of the transformation

On the transformed scale the picture changes (Figure 3, right). The skewness statistic drops from -0.816 to 0.068 , and the Shapiro–Wilk test gives $W = 0.975$ with $p = 0.921$, compared with $p = 0.093$ before. The one-sided t -test on the transformed values gives $t = 3.133$ and $p = 0.0037$, a much stronger result than on the original scale, now obtained under an assumption that the data support.

5.5 Is the transformation necessary?

The likelihood-ratio test of $H_0: \lambda = 1$ gives $\Lambda = 3.916$ and $p = 0.048$: at the 5% level the data reject the hypothesis that no transformation is needed. Figure 4 shows the profile log-likelihood of λ , obtained by maximising over μ and σ^2 for each value of λ . Its maximum is at $\hat{\lambda} = 1.305$, and the value $\lambda = 1$ lies just outside the set of values whose profile log-likelihood is within $\chi^2_{1,0.95}/2$ of the maximum, an approximate 95% confidence interval that runs from 1.005 to 1.690. The evidence against $\lambda = 1$ is therefore real but moderate, which is what a sample of fifteen observations can offer.

5.6 Conclusion on the original scale

With $\gamma = 0.01$, equation (11) gives

$$\hat{q}_{0.01} = \psi^{-1}\left(4.570 + t_{0.01,14}\sqrt{29.785/15}, 1.305\right) = 0.790. \quad (14)$$

Since even this lower quantile is positive, the analysis concludes, with more confidence than the original t -test, that cross-fertilised plants grow taller on average.

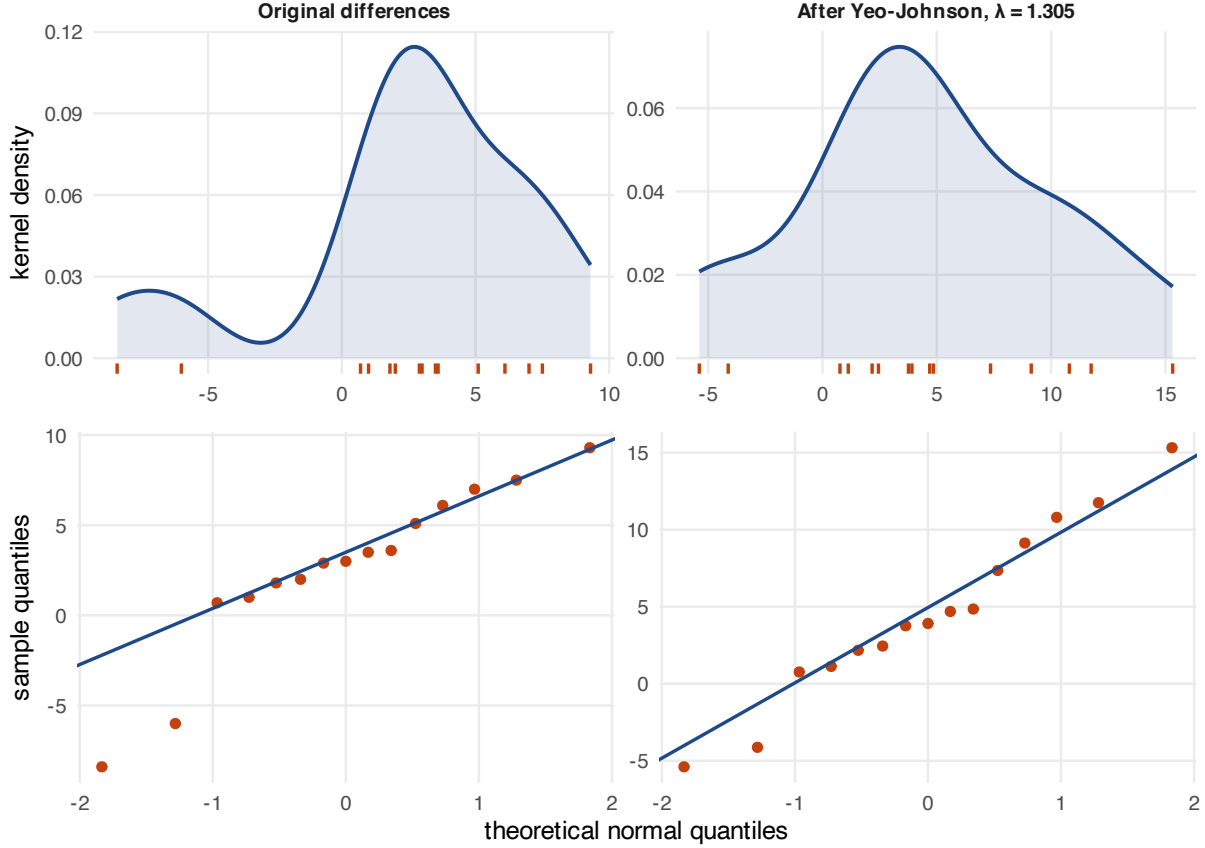


Figure 3: Darwin’s differences before (left) and after (right) the Yeo–Johnson transformation with $\hat{\lambda} = 1.305$. Top: kernel density estimates, with the observations marked below. Bottom: normal Q–Q plots. The long left tail of the original data disappears after the transformation.

5.7 Comparison with the published results

Table 2 compares the results of the scripts with those reported by Yeo and Johnson [11]. The estimates, the covariance matrix and the quantile agree to the precision of the paper; the small differences in the off-diagonal and variance terms of the covariance matrix come from the numerical Hessian.

Two differences deserve a comment. The likelihood-ratio statistic of the scripts evaluates the model with $\lambda = 1$ at the unbiased sample variance; with the maximum-likelihood variance, the one implied by the profile likelihood of Figure 4, the statistic becomes 3.88, essentially the published value, and both versions lead to the same conclusion at the 5% level. The p -value of the Shapiro–Wilk test on the transformed data (0.921 here, 0.887 in the paper) depends on the approximation used by the software, while the statistic W itself agrees. Skewness values are not compared directly because different sample-skewness conventions are in use; both analyses show the same change, from a clear left skew to a value close to zero.

6 Implementation

The code is organised in four R scripts. `functions.R` contains everything that is reused, and each of the other three scripts loads it and carries out one part of the study (Table 3).

The functions follow a defensive style: every transformation checks that its input is a numeric vector and its parameter a numeric scalar, and the Box–Cox function refuses non-positive data with an explicit error. The special cases $\lambda = 0$ and $\lambda = 2$ of (5) are handled as separate branches

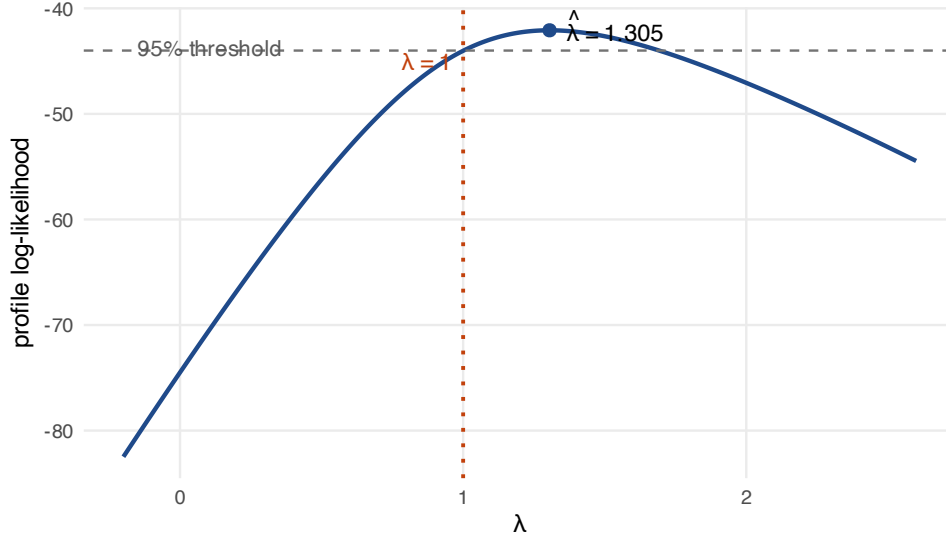


Figure 4: Profile log-likelihood of λ for Darwin’s data. The dashed line marks the 95% threshold, $\chi^2_{1,0.95}/2 \approx 1.92$ below the maximum; the dotted line marks $\lambda = 1$, the identity transformation, which falls just below the threshold.

Quantity	This project	Yeo and Johnson (2000)
Paired t -statistic, original data	2.142 ($p = 0.025$)	2.142 ($p = 0.025$)
$\hat{\lambda}$	1.305	1.305
$\hat{\mu}$	4.570	4.570
$\hat{\sigma}^2$	29.785	29.786
Covariance entries $(\lambda, \lambda), (\lambda, \mu), (\mu, \mu)$	0.434, 3.270, 54.445	0.434, 3.270, 54.451
Covariance entries $(\lambda, \sigma^2), (\mu, \sigma^2), (\sigma^2, \sigma^2)$	26.278, 198.175, 3366.98	26.282, 198.227, 3367.61
Shapiro–Wilk W , transformed data	0.975	0.975
Likelihood-ratio statistic for $\lambda = 1$	3.916 ($p = 0.048$)	3.873 ($p = 0.050$)
Estimated 0.01-quantile of the mean difference	0.790	0.790

Table 2: Results of the R implementation compared with the values published in the original paper.

rather than through numerical limits, and the transformations are vectorised, so a whole sample is transformed in one call. Optimisation uses the base R function `optim` with the L-BFGS-B method in all three parts, with bounds on λ that keep the search in a meaningful range, and `hessian = TRUE` in the final joint fit. Random sampling is seeded, so that the experiments on the mixture can be repeated exactly. The core functions are listed in Appendix A.

7 Conclusions

The project followed the Yeo–Johnson transformation from its construction to its use in a real inferential problem. The experiment on a skewed mixture showed why a new family was needed: the signed power transformation, at its best parameter, still split the data into two peaks, while the Yeo–Johnson transformation, with its two coupled branches, produced a single, nearly normal distribution. On Darwin’s data, the likelihood-based treatment of λ , which includes the Jacobian of the transformation, gave estimates, a covariance matrix and a quantile that match the published ones, and turned a borderline t -test on doubtful data into a clear conclusion supported by a normality check.

Two points summarise what the analysis adds to a plain significance test. First, the transformation parameter can itself be estimated and tested, so the need for a transformation is a

File	Content
<code>functions.R</code>	Box–Cox, signed power and Yeo–Johnson transformations and the inverse Yeo–Johnson transformation; the mixture density; numerical tests of continuity, monotonicity and convexity; the negative log-likelihood (9); the skewness statistic; the quantile estimate (11).
<code>1st_part.R</code>	Sampling from the mixture, choice of the signed power parameter with L-BFGS-B, density plot.
<code>2nd_part.R</code>	Geometry of the Yeo–Johnson family, Box–Cox comparison, transformation of the mixture.
<code>3rd_part.R</code>	Darwin’s data: t -test, normality diagnostics, two-stage maximum likelihood with Hessian, transformed-scale tests, likelihood-ratio test, quantile estimate.

Table 3: Organisation of the code.

statistical question with a likelihood-ratio answer rather than a matter of taste. Second, results obtained on the transformed scale can be brought back to the original one through the inverse transformation, as the quantile of the mean difference shows. Natural extensions of the work are the use of the transformation in regression models, where λ is estimated jointly with the coefficients, and robust versions of the estimator that limit the influence of outlying observations on $\hat{\lambda}$.

References

- [1] P. J. Bickel and K. A. Doksum. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311, 1981.
- [2] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2):211–252, 1964.
- [3] R. H. Byrd, P. Lu, J. Nocedal and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.
- [4] R. J. Carroll and D. Ruppert. Prediction and tolerance intervals with transformation and/or weighting. *Technometrics*, 33(2):197–210, 1991.
- [5] C. Darwin. *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom*. John Murray, London, 1876.
- [6] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.
- [7] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [9] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.
- [10] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, 2016.
- [11] I.-K. Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

A Core functions

The Yeo–Johnson transformation and the negative log-likelihood, from `functions.R` (input checks omitted).

```
yeo_johnson <- function(x, lambda) {
  y <- numeric(length(x))
  pos_values <- x >= 0
  neg_values <- x < 0
  y[neg_values] <- if (lambda == 2) {
    -log(-x[neg_values] + 1)
  } else {
    -((-x[neg_values] + 1)^(2 - lambda) - 1) / (2 - lambda)
  }
  y[pos_values] <- if (lambda == 0) {
    log(x[pos_values] + 1)
  } else {
    ((x[pos_values] + 1)^lambda - 1) / lambda
  }
  return(y)
}

log_likelihood <- function(lambda, mu, sigma2, x) {
  transformed_x <- yeo_johnson(x, lambda)
  n <- length(x)
  log_likelihood_value <- -(n / 2) * log(2 * pi) - (n / 2) * log(sigma2) -
    (1 / (2 * sigma2)) * sum((transformed_x - mu)^2) +
    (lambda - 1) * sum(sign(x) * log(abs(x) + 1))
  return(-log_likelihood_value) # negative, for minimisation
}
```

The two-stage estimation of Section 4, from `3rd_part.R`:

```
objective_function <- function(lambda, data) {
  transformed_data <- yeo_johnson(data, lambda)
  log_likelihood(lambda, mean(transformed_data), var(transformed_data), data)
}

best_lambda <- optim(par = 0, fn = objective_function, data = data,
  method = "L-BFGS-B", lower = -2, upper = 2)$par
transformed_data <- yeo_johnson(data, best_lambda)
optim_result <- optim(par = c(best_lambda, mean(transformed_data), var(transformed_data)),
  fn = function(p) log_likelihood(p[1], p[2], p[3], data),
  hessian = TRUE, method = "L-BFGS-B")
covariance_matrix <- solve(optim_result$hessian) * 15 # n times the inverse Hessian
```