

# The powerLaw package: Comparing distributions

Colin S. Gillespie

Last updated: January 10, 2019

The `powerLaw` package provides an easy to use interface for fitting and visualising heavy tailed distributions, including power-laws. This vignette provides examples of comparing competing distributions.

## 1 Comparing distributions

This short vignette aims to provide some guidance when comparing distributions using Vuong's test statistic. The hypothesis being tested is

$H_0$  : Both distributions are equally far from the true distribution

and

$H_1$  : One of the test distributions is closer to the true distribution.

To perform this test we use the `compare_distributions` function<sup>1</sup> and examine the `p_two_sided` value.

## 2 Example: Simulated data 1

First let's generate some data from a power-law distribution

```
library("powerLaw")
set.seed(1)
x = rpldis(1000, xmin=2, alpha=3)
```

and fit a discrete power-law distribution

```
m1 = displ$new(x)
m1$setPars(estimate_pars(m1))
```

The estimated values of  $x_{\min}$  and  $\alpha$  are 2 and 2.97, respectively. As an alternative distribution, we will fit a discrete Poisson distribution<sup>2</sup>

---

<sup>1</sup>The `compare_distributions` function also returns a one sided  $p$ -value. Essentially, the one sided  $p$ -value is testing whether the first model is better than the second, i.e. a **one** sided test.

<sup>2</sup>When comparing distributions, each model must have the same  $x_{\min}$  value. In this example, both models have  $x_{\min} = 2$ .

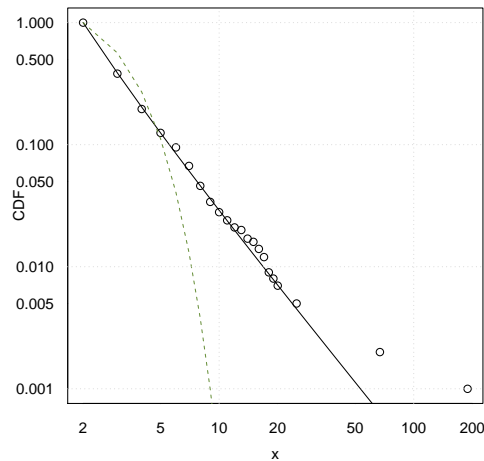


Figure 1: Plot of the simulated data CDF, with power law and Poisson lines of best fit.

```
m2 = dispois$new(x)
m2$setPars(estimate_pars(m2))
```

Plotting both models

```
plot(m2, ylab="CDF")
lines(m1)
lines(m2, col=2, lty=2)
```

suggests that the power-law model gives a better fit (figure 1). Investigating this formally

```
comp = compare_distributions(m1, m2)
comp$p_two_sided

## [1] 0.05141726
```

means we can reject  $H_0$  since  $p = 0.05142$  and conclude that one model is closer to the true distribution.

### One or two-sided $p$ -value

The two-sided  $p$ -value does not depend on the order of the model comparison

```
compare_distributions(m1, m2)$p_two_sided

## [1] 0.05141726

compare_distributions(m2, m1)$p_two_sided

## [1] 0.05141726
```

However, the one-sided  $p$ -value is order dependent

```
## We only care if m1 is better than m2
## m1 is clearly better
compare_distributions(m1, m2)$p_one_sided

## [1] 0.02570863

## m2 isn't better than m1
compare_distributions(m2, m1)$p_one_sided

## [1] 0.9742914
```

### 3 Example: Moby Dick data set

This time we will look at the Moby Dick data set

```
data("moby")
```

Again we fit a power law

```
m1 = displ$new(moby)
m1$setXmin(estimate_xmin(m1))
```

and a log-normal model<sup>3</sup>

```
m2 = dislnorm$new(moby)
m2$setXmin(m1$getXmin())
m2$setPars(estimate_pars(m2))
```

Plotting the CDFs

```
plot(m2, ylab="CDF")
lines(m1)
lines(m2, col=2, lty=2)
```

suggests that both models perform equally well (figure 2). The formal hypothesis test

```
comp = compare_distributions(m1, m2)
```

gives a  $p$ -value and test statistic of

```
comp$p_two_sided
## [1] 0.6773366

comp$test_statistic
## [1] 0.4161003
```

which means we can not reject  $H_0$ . The  $p$ -value and test statistic are similar to the values found in table 6.3 of [Clauset et al. \(2009\)](#).

---

<sup>3</sup>In order to compare distributions,  $x_{\min}$  must be equal for both distributions.

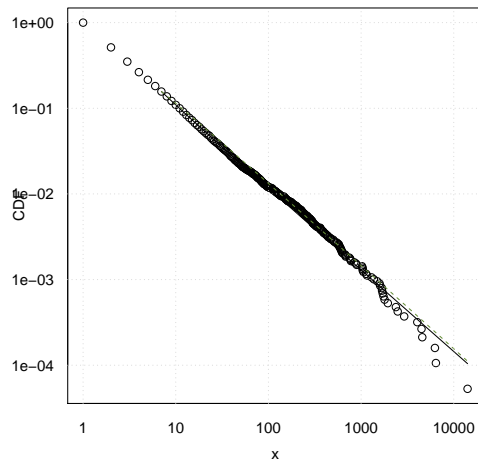


Figure 2: The Moby Dick data set with power law and log normal lines of best fit.

## References

- A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.