Fitting degree distribution

```r
library(dplyr)
library(poweRlaw)
source("order_fit_distributions.R")
set.seed(156)
```

Fitting degree distribution

Using only 2002 as an example, we can use the "poweRlaw" package to fit different models to the in-degree distribution. This package is based on the Newman paper "Power-law distributions in empirical data" (2009). The purpose of that paper and this section is to empirically deduce the underlying models without assuming it is a pwoer law.

poweRlaw deals with 4 types of discrete model: power law, lognormal, exponential and poisson. We can use poweRlaw to fit these models to the data, estimating their parameters and minimum value of x for which they apply.

Finding p values

Using a bootstrapping method to obtain a p value for distributions

- The paper says that for a given accuracy in the p value the number of simulations required scales $\frac{1}{4}\eta^{-1/2}$.
- The paper also approximates a p value cut off of about 0.1 to disregard the hypothesis of the tested distribution.
- To get an accuracy of 0.01 we need 2500 simulations, this is would take far too long (~6 hours for power law, and ~ 27hours for lognormal)

We find that over 2500 bootstrapped simulations the goodness of fit is never more extreme than our data. This means that **both distributions yield a p value of 0**. I think this is caused by the deviation from the fit towards the tail. Perhaps I should rerun the bootstrapping removing the tail?

To find which is a better fit

- one sided p value is the upper limit on getting that small a log-likelihood ratio if the first distribution (m_pl) is true.
- two sided p value is the probability of getting a log-likelihood ratio which deviates that much from zero in either direction if the two distributions are equally good.
- Test statistic is the sample average of the log-likelihood ratio standardised by an estimated standard deviation.

In the above we see equally low order p values reafirming that it is hard to strongly link the data to a particular distribution. The test statistic states that the log-normal model fits better and looking at the plot we see that it is initially in favour of the power-law distribution but becomes increasingly closer to the log-normal towards the tail.

This matches what we see in the tail of the distributions, after x = 150-200 there is a strange dip to the frequencies and log-normal follows this more closely, however during the linear section power law is a better fit.

```r
x1978$comparisons[[1]][1:3]
```

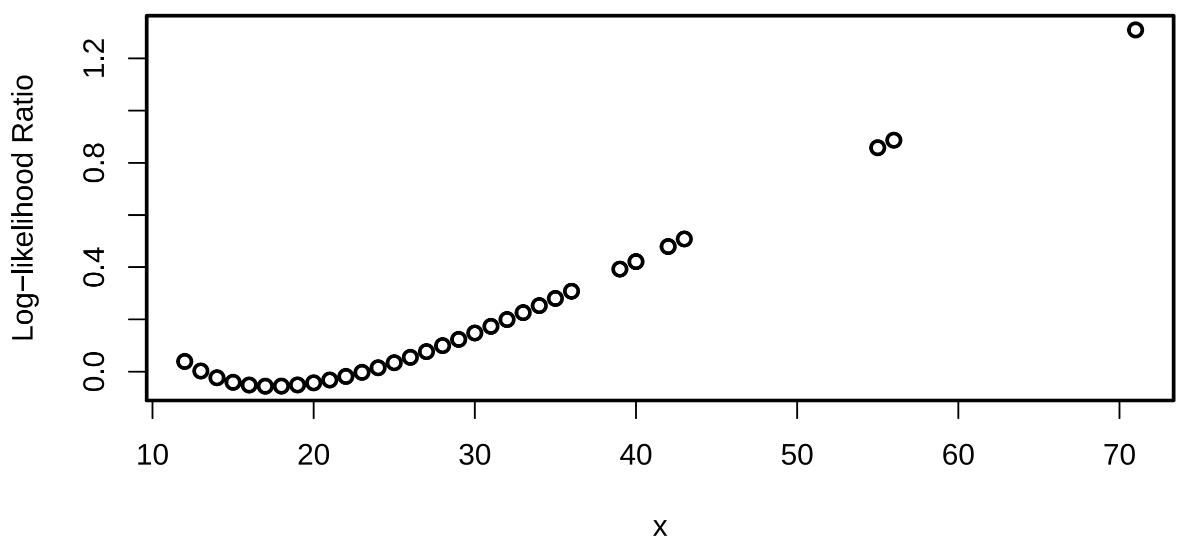
```
## $test_statistic
## [1] -1.66488
```

```
##
## $p_one_sided
## [1] 0.04796831
##
## $p_two_sided
## [1] 0.09593661
```

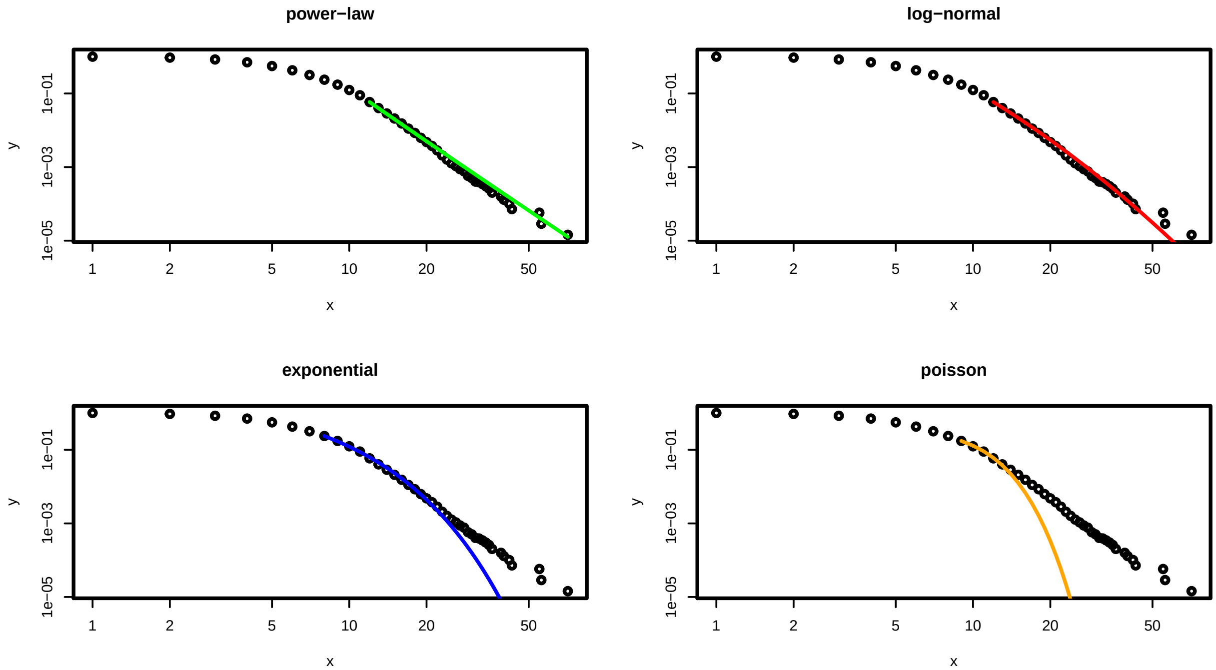
```
x1978$comparisons[[2]][1:3]
```

```
## $test_statistic
## [1] 1.66488
##
## $p_one_sided
## [1] 0.9520317
##
## $p_two_sided
## [1] 0.09593661
```

```
x1978$plots[[2]]
```




```
x1978$plots[[3]]
```



The above plot shows power law and exponential functions as a good fit whereas the exponential and poisson distributions are not. While both the one sided p values are not unsubstantial and the two sided p value (belonging to both distributions) is also statistically significant the p value for power law is very high. As a theme the log-likelihood ratio varies by at first being slightly in favour of power law before briefly switching and becoming more extreme in the tail.

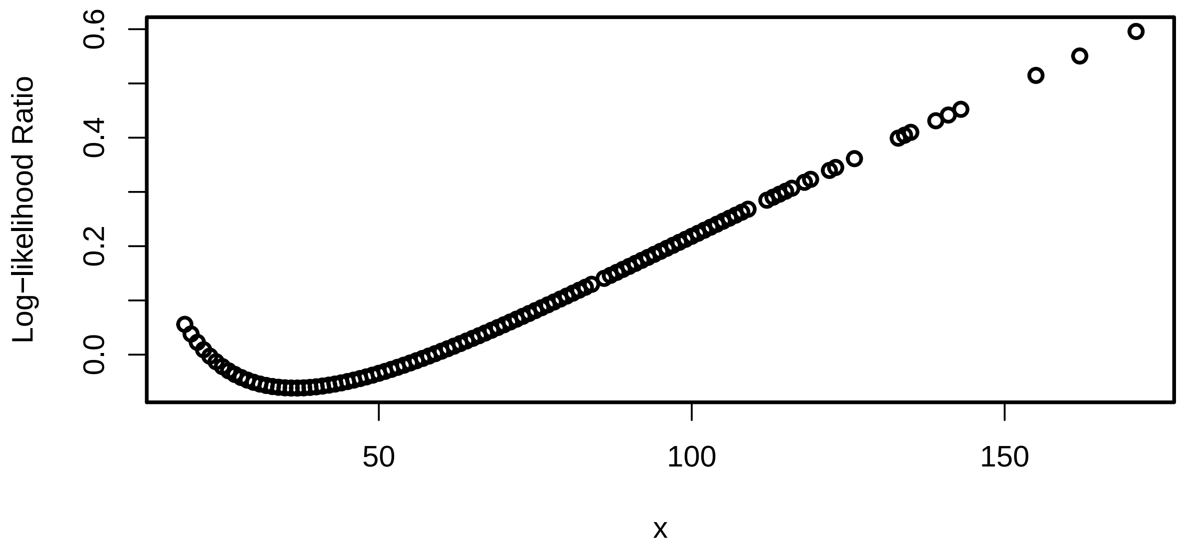
```
x1992$comparisons[[1]][1:3]
```

```
## $test_statistic
## [1] -3.635451
##
## $p_one_sided
## [1] 0.0001387472
##
## $p_two_sided
## [1] 0.0002774944
```

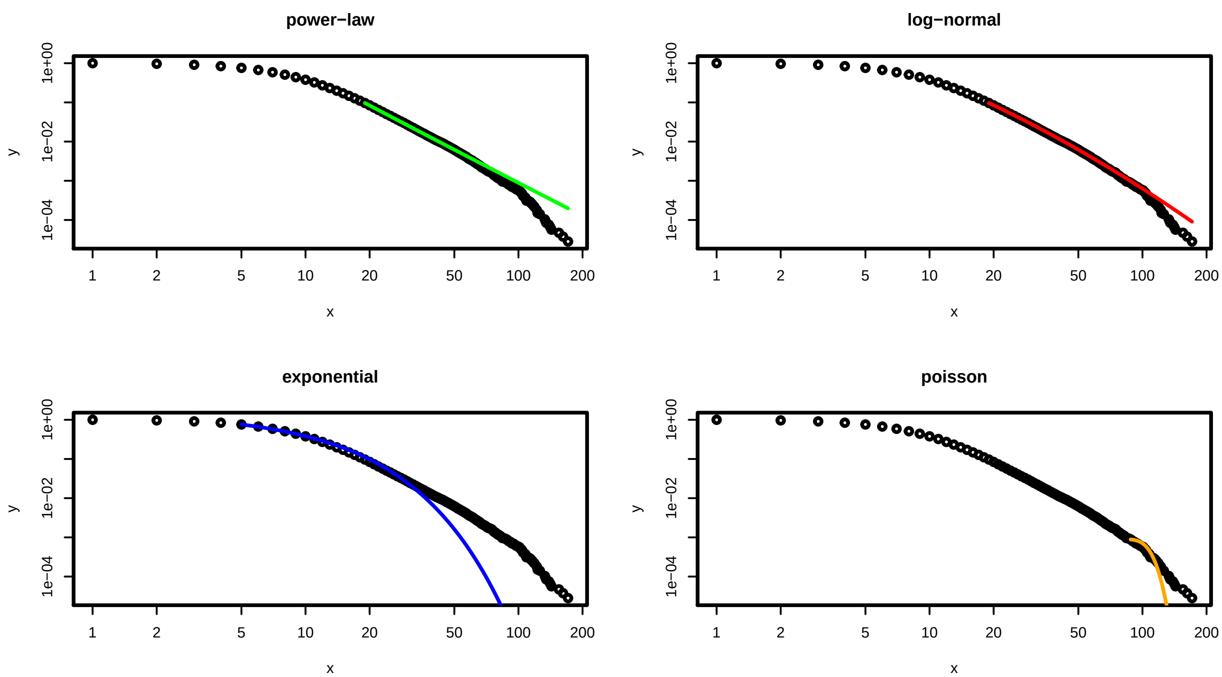
```
x1992$comparisons[[2]][1:3]
```

```
## $test_statistic
## [1] 3.635451
##
## $p_one_sided
## [1] 0.9998613
##
## $p_two_sided
## [1] 0.0002774944
```

```
x1992$plots[[2]]
```




```
x1992$plots[[3]]
```




```
x2002$comparisons[[1]][1:3]
```

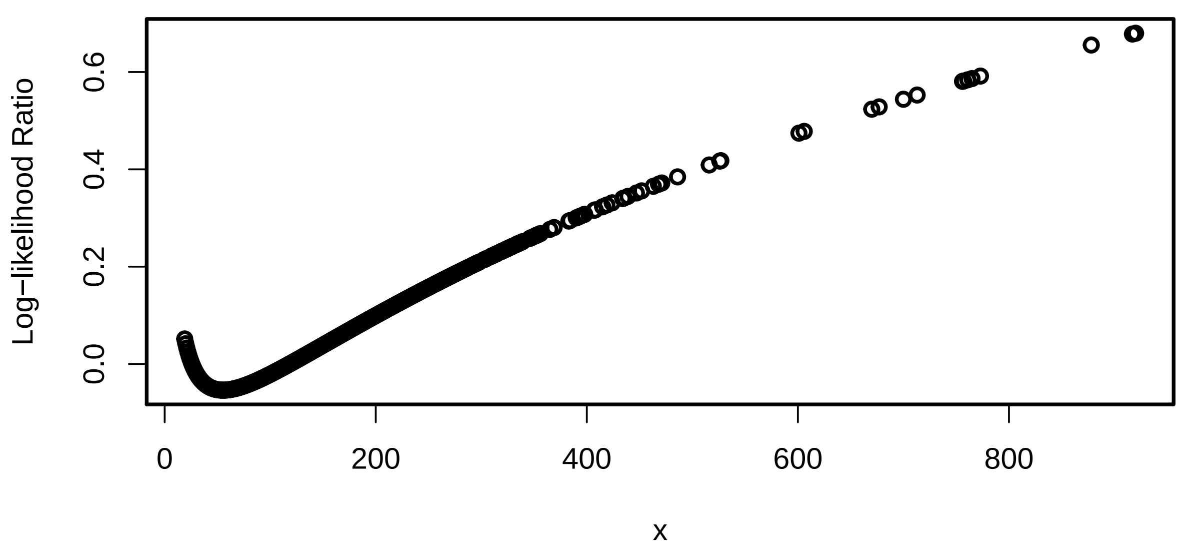
```
## $test_statistic
## [1] -6.342428
##
## $p_one_sided
```

```
## [1] 1.130857e-10
##
## $p_two_sided
## [1] 2.261714e-10
```

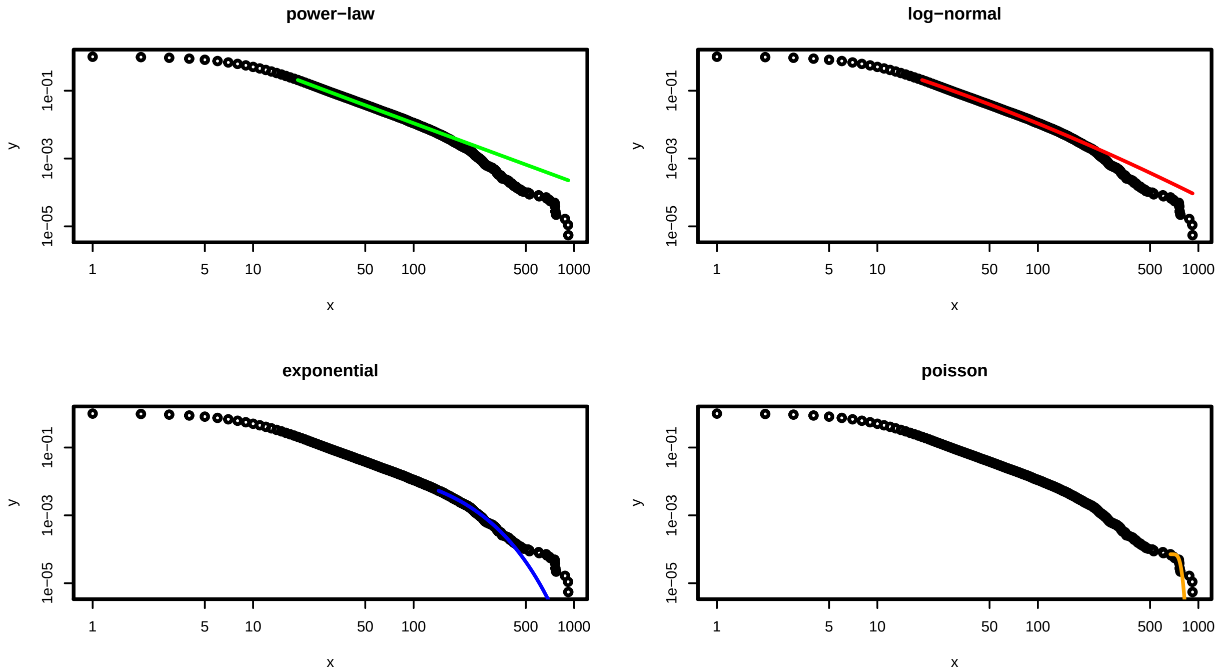
```
x2002$comparisons[[2]][1:3]
```

```
## $test_statistic
## [1] 6.342428
##
## $p_one_sided
## [1] 1
##
## $p_two_sided
## [1] 2.261715e-10
```

```
x2002$plots[[2]]
```




```
x2002$plots[[3]]
```



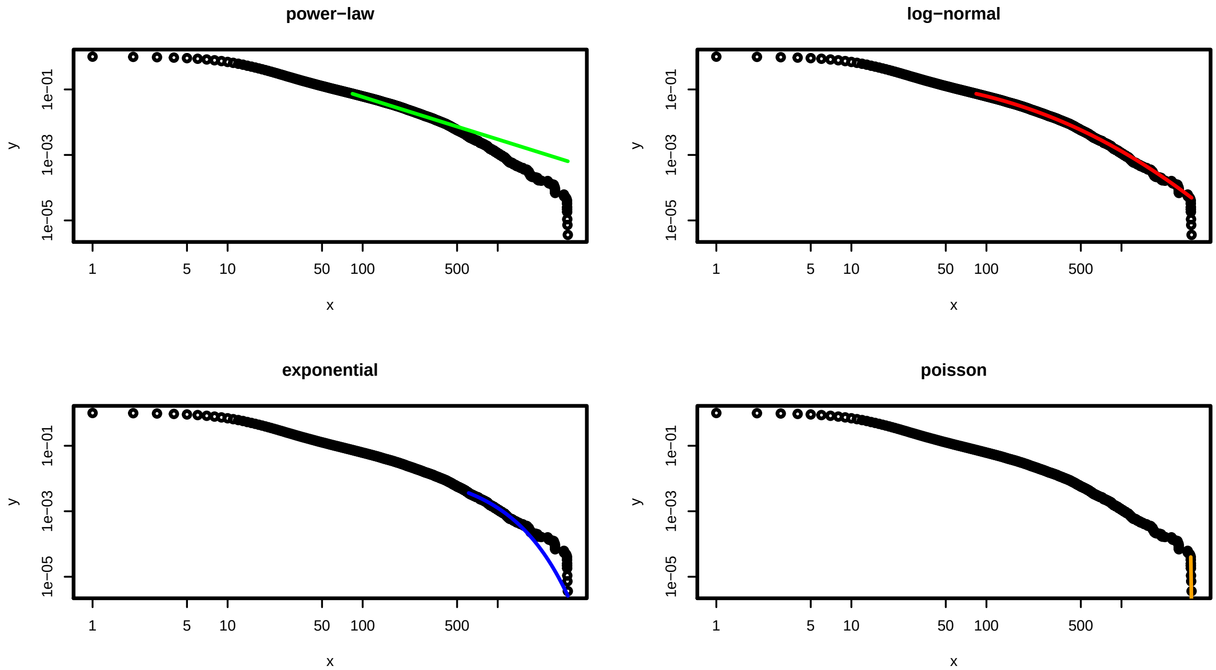

```
x2012$comparisons[[1]][1:3]
```

```
## $test_statistic
## [1] -19.5905
##
## $p_one_sided
## [1] 9.318451e-86
##
## $p_two_sided
## [1] 1.86369e-85
```

```
x2012$comparisons[[2]][1:3]
```

```
## $test_statistic
## [1] 19.5905
##
## $p_one_sided
## [1] 1
##
## $p_two_sided
## [1] 0
```

```
x2012$plots[[3]]
```




```
x2012$plots[[2]]
```