We have, as noted throughout this paper, tried to make the statistical methodology as accessible as possible by implementing it in an R package. In this section we will demonstrate the use of the package with data from various surveys.

## 1.1 Fitting a zeta distribution

We start with the data from Roux et al. [1] who conducted a footwear survey, where shoes from 776 individuals were examined for glass. This data set is built into the package and can be accessed from the `Psurveys` object. That is, we can type:

```
> data("Psurveys")
> roux = Psurveys$roux
```

The package includes a special printing function that summarises the data for reading rather than displaying it in the way it is stored. R prints the values of objects (or variables) simply by typing their name. For example

```
> roux

Number of Groups

  n    rn
---  ----
  0   754
  1     9
  2     8
  3     4
  4     1
Roux C, Kirk R, Benson S, Van Haren T, Petterd C (2001).
"Glass particles in footwear of members of the public in
south-eastern Australia-a survey." _Forensic Science
International_, *116*(2), 149-156.
doi:10.1016/S0379-0738(00)00355-8
<https://doi.org/10.1016/S0379-0738%2800%2900355-8>.
```

It is very simple to fit a zeta distribution to this data set. We do this using the `fitDist` function.

```
> fit = fitDist(roux)
```

There is nothing special about the variable name `fit`. It could be anything, for example `x`, or `blah`. However we choose the variable name `fit` because it is easy to remember that it is a *fitted* object. The package includes specialised functions for both printing and plotting the fitted object. The `print` method

displays an estimate of the shape parameter $\alpha$, an estimate of the standard deviation—the standard error—of the estimate of $\alpha$ ($\widehat{\text{sd}}(\hat{\alpha}) = \text{se}(\hat{\alpha})$). **Note:** it is important to understand that the value of the shape parameter that is displayed, and the value that is stored in the fitted object differ by 1. That is, $\alpha$ is shown, and $s = \alpha - 1$ is stored. This difference only has consequences if the fitted value is being used in conjunction with other functions.

### 1.1.1 Using the fitted distribution to estimate $P$ terms

The `print` method also displays the first 10 fitted probabilities from the model by default.

```
> fit

The estimated shape parameter is 4.9544
The standard error of shape parameter is 0.2366
------
NOTE: The shape parameter is reported so that it is consistent
with Coulson et al. However, the value returned is actually s'
= shape - 1 to be consistent with the VGAM parameterisation,
which is used for computation. This has flow on effects, for
example in confInt. This will be changed at some point.
------

The first  10 fitted values are:
           P0           P1           P2           P3           P4
9.631547e-01 3.106447e-02 4.167082e-03 1.001917e-03 3.316637e-04
           P5           P6           P7           P8           P9
1.344002e-04 6.262053e-05 3.231467e-05 1.802885e-05 1.069709e-05
```

This information is probably sufficient for most casework. However, the package has a function, `probfun`, that returns a bespoke function that can calculate any probability term. This function is applied a fitted object. For example

```
> P = probfun(fit)
```

Again, `P` is just a variable name and we could have used anything. We have chosen `P` because this probability function returns $P$ terms. To use it, we type only need to provide the value of $k$, and the function will return $P_k$. For example

```
> P(5)

           P5
0.0001344002
```

## 1.2 Fitting a zero-inflated zeta distribution

We can also easily fit a zero-inflated zeta model using the `fitZIDist` function[1]. As before, we can choose a variable name to store the results in.

```
> fit.zi = fitZIDist(roux)
> fit.zi

The estimated mixing parameter, pi, is 0.8465
The estimated shape parameter is 2.8846
------
NOTE: The shape parameter is reported so that it is consistent
with Coulson et al. However, the value returned is actually s'
= shape - 1 to be consistent with the VGAM parameterisation,
which is used for computation. This has flow on effects, for
example in confInt. This will be changed at some point.
------

The first  10 fitted values are:
          P0           P1           P2           P3           P4
0.9716490911 0.0169404164 0.0052597614 0.0022938450 0.0012050764
          P5           P6           P7           P8           P9
0.0007122067 0.0004565511 0.0003106019 0.0002211302 0.0001631754
```

In the example above we fit a zero-inflated model to Roux et al.'s data, and print out the resulting fit. We get, as with the zeta model, estimates of the parameters and a default set of fitted values. The output is interesting in that we can see (from the value of $\hat{\pi}$) that the *zero* part of the zero-inflated model is picking up about 85% of the zeros. It is interesting to compare the estimates from the raw frequencies, the zeta model, and those of the ZIZ model. The estimates are shown in Table 1.

| $k$ | $P_k^{raw}$ | $P_k^{zeta}$ | $P_k^{ZIZ}$ |
|---|---|---|---|
| 0 | 0.9716 | 0.9632 | 0.9716 |
| 1 | 0.0116 | 0.0311 | 0.0169 |
| 2 | 0.0103 | 0.0042 | 0.0053 |
| 3 | 0.0052 | 0.0010 | 0.0023 |
| 4 | 0.0013 | 0.0003 | 0.0012 |
| 5 | 0.0000 | 0.0001 | 0.0007 |

Table 1: Estimated probability that $k$ groups of glass would be found in shoes of a random member of the population based on the data of [1], the raw frequencies, and those produced from the zeta and ZIZ models respectively.

We can see from Table 1 that we now have a non-zero estimate for $P_5$,

[1]Functions with mixed case names are often annoying. For that reason, the package also allows `fitZIdist` and `fitzidist`.

but this comes at a small cost, namely smaller probabilities for the preceding terms $P_0$–$P_4$. This is not necessarily a negative. The survey data is dominated by zeros. However, we think it likely that the raw sample estimates (for $P_0$–$P_4$) are over-estimates. The model reduces the estimated value, which is in line with our thinking. Interestingly, the effect of including the zero-inflation factor is to increase nearly all of the probabilities, with the exception of $P_1$. A natural question to ask is "Which model is correct?" The answer, unhelpfully, is "Neither." This is because these are simply models. They can still help us without us having to believe that they are true.

## 1.3 Confidence intervals for the parameter estimates

The `fitPS` package provides a `confint` method for the fitted value. The method returns both a Wald confidence interval and profile likelihood interval. The two intervals are returned as elements of a `list` named `wald` and `prof` respectively.

```
> ci = confint(fit)
> ci$wald

    2.5%    97.5%
3.490761 4.418099

> ci$prof

    2.5%    97.5%
3.520495 4.451277
```

Readers may notice neither of these intervals contain the estimated value shown in the previous output. However, this simply because they are confidence intervals on $s'$ and not $s$. This can be remedied by adding one to each interval:

```
> ci$wald + 1

    2.5%    97.5%
4.490761 5.418099

> ci$prof + 1

    2.5%    97.5%
4.520495 5.451277
```

The reason for not *correcting* these intervals is that the method mostly exists to feed into other parts of the package, especially the `plot` method.

### 1.3.1 Bootstrapped and profile likelihood confidence regions for the zero-inflated zeta

The package includes the facility to compute both bootstrapped and profile likelihood confidence regions for the parameters of the zero-inflated zeta distribtion. It does, also, in fact compute bootstrapped confidence intervals for the zeta distribution, but we do not do demonstrate this functionality here. The `confint` function returns a confidence region if the fitted object contains information from a zero-inflated zeta fit. As an example, we will first compute profile likelihood confidence regions for the Roux et al. [1] data. To do this we use the fitted object we previously created, `fit.zi`, and, although not required, we supply a set of two levels so that we can compute both an 80% and a 95% confidence region. `confit` returns a list of confidence regions—one for each level—each of which are simply a set of $x$ and $y$ coordinates corresponding to the appropriate contour line. We can use this information for plotting. The code to produce Figure 1 is given below.

```
> cr = confint(fit.zi, level = c(0.80, 0.95))
> plot(cr[["0.95"]], type = "l")
> polygon(cr[["0.8"]], border = "red")
> legend("topright", lty = 1, lwd = 2, col = c("red", "black"),
+        legend = c("80%", "95%"), bty = "n")
```

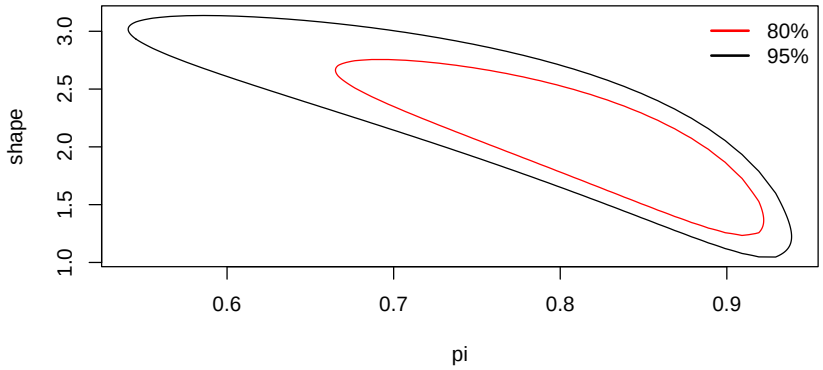


Figure 1: 80% and 95% confidence regions for the parameters of a zero-inflated zeta distribution fitted to the Roux et al. [1] data.

A bootstrapped confidence region can be computed using the `bootCI` function. The `bootCI` function includes the facility to plot the resulting confidence region(s) as and hide or display the function's progress. The latter is important

because this procedure is numerically intensive, and even with utilising parallel processing, can be quite slow. The code below produces Figure 2

```
> bcr = bootCI(roux, model = "ziz", plot = TRUE, silent = TRUE)
```

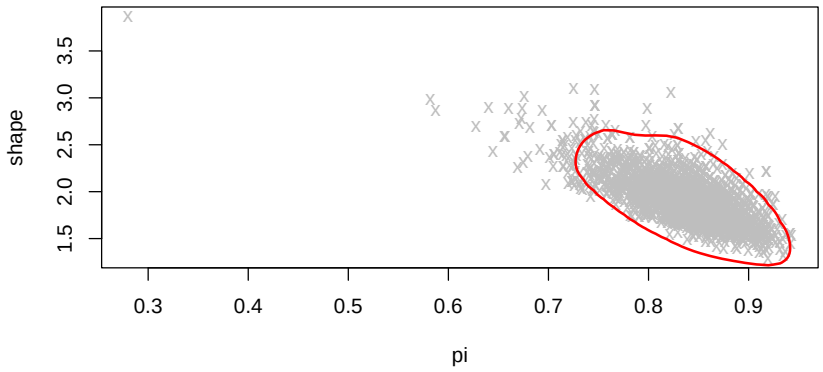


Figure 2: A 95% bootstrapped confidence region for the parameters of a zero-inflated zeta distribution fitted to the Roux et al. [1] data.

## 1.4 Comparing two surveys

We can use the methodology that has been demonstrated so far to compare surveys. One reason for comparing surveys is to explore the hypothesis that there is no difference in the underlying "true[2]" value of $\alpha$. If there is insufficient evidence to reject this hypothesis, then one may feel justified in combining data from two surveys. In the first instance we will take an ad-hoc approach, and then treat this problem more formally. In our ad-hoc approach we will compare confidence intervals for two-surveys. If these confidence intervals overlap, then we might conclude that there is insufficient evidence in the data to suggest that the estimates of $\alpha$ are different. We will illustrate this with the surveys conducted by Lau et al. [2] and Jackson et al. [3]. Lau et al. [2] surveyed the clothing of 213 Canadian high school students and observed two sets of clothing with one fragment on each. Similarly, Jackson et al. [3] surveyed 232 "randomly" selected members of the population of New South Wales in Australia. We place "randomly" in quotes because this was not a true random sample, but rather a

[2]Readers may be aware that one of the authors is fundementally Bayesian at heart, and the so the concept of the true value of a population parameter is antithetical to the school of thought. We, however, proceed on the basis that the Frequenist school of thought is not incorrect, but just differs in interpretation.

convenience sample. That being said, it is unlikely that using a truly random mechanism would have significantly changed the results. Jackson et al. [3] found a single fragment of glass on each of six people. The data from each of these two surveys is summarised in Table 2.

| | $r_n$ | |
|---|---|---|
| $n$ | Lau et al. | Jackson et al. |
| 0 | 211 | 224 |
| 1 | 2 | 6 |

Table 2: Survey results from Lau et al. [2] and Jackson et al. [3].

Visual inspection of these surveys would suggest that they are fairly similar. We can fit a zeta distribution to each survey, and then compute a confidence interval for each survey. Again, these data sets are included in the `fitPS` package.

```
> lau = Psurveys$lau
> jackson = Psurveys$jackson
> fit.lau = fitDist(lau)
> fit.jackson = fitDist(jackson)
> confint(fit.lau)$wald

    2.5%    97.5%
3.948422 7.823417

> confint(fit.jackson)$wald

    2.5%    97.5%
3.443639 5.623473
```

We can see from the output that there is substantial overlap between these two (Wald) confidence intervals. The results using profile likelihood intervals lead to the same conclusion but are not shown. We can test this more formally. Specifically, we wish to test the (null) hypothesis that

$$H_0 : \alpha_1 = \alpha_2 \text{ or equivalently } H_0 : \alpha_1 - \alpha_2 = 0,$$

where $\alpha_1$ is the true value of $\alpha$ for the the Lau et al. data, and $\alpha_2$ is the true value of $\alpha$ for the the Jackson et al. data. We choose a two-tailed alternative, meaning we are not concerned about the sign of any difference, but simply the magnitude of the difference. That is,

$$H_1 : \alpha_1 \neq \alpha_2 \text{ or equivalently } H_1 : \alpha_1 - \alpha_2 \neq 0.$$

We test this hypothesis by constructing a test statistic, then computing a $P$-value under the assumption that the null hypothesis is true. We are interested in the difference between the two population values of $\alpha$. We estimate this by

computing the difference in the sample estimates. That is, our estimate of $\alpha_1 - \alpha_2$, is given by $\hat{\alpha}_1 - \hat{\alpha}_2$, where $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are the maximum likelihood estimates based on the survey data. We scale this difference by the estimated standard deviation in the difference, that is, by the standard error of the difference, $se(\hat{\alpha}_1 - \hat{\alpha}_2)$. We estimate this—to keep the statistical theory to a minimum—as the sum of the square root of the two estimated variances, i.e.

$$se(\hat{\alpha}_1 - \hat{\alpha}_2) = \sqrt{\hat{V}(\hat{\alpha}_1) + \hat{V}(\hat{\alpha}_2)}.$$

Our test statistic is then

$$Z_0 = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{se(\hat{\alpha}_1 - \hat{\alpha}_2)}.$$

It can be shown that this test statistic follows an approximate normal distribution under the null hypothesis. That means we can compute our $P$-value by evaluating

$$\begin{aligned} P &= \Pr(Z > |Z_0|) \\ &= 2(1 - \Pr(Z < |Z_0|)). \end{aligned}$$

All this theory has been put in a function called `compareSurveys`

```
> compareSurveys(lau, jackson)


Two-sided Wald test

data:  lau and jackson
z = 1.1923, p-value = 0.2331
alternative hypothesis: true difference in shape parameters is not equal to 0
sample estimates:
    Shape of lau Shape of jackson
        5.885919         4.533556
```

We can see that the $P$-value is 0.23 (2 d.p.). This is significantly larger than either 0.05, or 0.01. Based on this we would conclude that there is insufficient evidence to reject the null hypothesis of a common value of $\alpha$, and therefore it may be sensible to combine data from these two surveys. We could have also used the theory of likelihood ratio tests [4] to test this hypothesis, but that is beyond the scope of this article. We note, however, that the `fitPS` package contains a function called `compareSurveysLRT` which can compare two *or more* surveys simultaneously using a likelihood ratio test.

# References

[1] C. Roux, R. Kirk, S. Benson, T. Van Haren, and C. I. Petterd. Glass particles in footwear of members of the public in south-eastern australia—a

survey. *Forensic Science International*, 116(2):149–156, 2001. doi: https://doi.org/10.1016/S0379-0738(00)00355-8.

[2] L. Lau, A. D. Beveridge, B. C. Callowhill, N. Conners, K. Foster, R. J. Groves, K. N. Ohashi, A. M. Sumner, and H. Wong. The frequency of occurrence of paint and glass on the clothing of high school students. *Canadian Society of Forensic Science Journal*, 30(4):233–240, 1997. doi: 10.1080/00085030.1997.10757103.

[3] F. Jackson, P. Maynard, K. Cavanagh-Steer, T. Dusting, and C. Roux. A survey of glass found on the headwear and head hair of a random population vs. people working with glass. *Forensic Science International*, 226(1):125–131, 2013. doi: https://doi.org/10.1016/j.forsciint.2012.12.017.

[4] Wikipedia contributors. Likelihood-ratio test, 2024. URL `https://en.wikipedia.org/wiki/Likelihood-ratio_test`. [Online; accessed 8-January-2024].