# PF Package Vignette

David Siev, Christopher H. Tong

January 4, 2011

## Contents

# 1 Introduction

The `PF` package is a collection of functions related to estimating prevented fraction, $PF = 1 - RR$ , where $RR = \pi_2/\pi_1$ .

### Technical notes

*Optimization.* Unless otherwise stated, optimization is by the DUD algorithm (Ralston and Jennrich 1978).

*Level tested.* The help files indicate the level of testing undergone by each function. In some cases that is a subjective judgement, since most of these functions were originally tested in SPlus and have been ported to R more recently.

# 2 Score based methods

## 2.1 The score statistic

Confidence intervals for the risk ratio may be based on the score statistic (Koopman 1984; Miettinen and Nurminen 1985),

$$\frac{\hat{\pi}_2 - \rho_0 \hat{\pi}_1}{\sqrt{(\tilde{\pi}_2(1-\tilde{\pi}_2)/n_2\ ) + \rho_0^2\,(\tilde{\pi}_1(1-\tilde{\pi}_1)/n_1\ )}}$$

where hat indicates the MLE and tilde indicates the MLE under the restriction that $\rho = \rho_0$.

## 2.2 Asymptotic intervals

`RRsc()` estimates asymptotic confidence intervals for the risk ratio or prevented fraction based on the score statistic. Interval estimates are returned for three estimators. The score method was originally introduced by Koopman (1984). Gart and Nam's modification includes a skewness correction (Gart and Nam 1988). The method of Miettinen and Nurminen (1985) is a version made slightly more conservative than Koopman's by including a factor of $(N-1)/N$ .

```
> require(PF)
> RRsc(c(4,24,12,28))

PF
95% interval estimates

                PF     LL    UL
MN method    0.611 0.0251 0.857
score method 0.611 0.0328 0.855
skew corr    0.611 0.0380 0.876
```

Starting estimates for the algorithm are obtained by the modified Katz method (log method with 0.5 added to each cell). Both forms of the Katz estimate may be retrieved from the returned object using `RRsc()$estimate`.

## 2.3 Exact intervals

These methods give intervals based on the score statistic that are 'exact' in the sense of accounting for discreteness. The score statistic is used to select the $2 \times 2$ tables that would fall in the tail area, and the binomial probability is estimated over the tail area by taking the maximum over the nuisance parameter. The search over the tail area is made more efficient by the Berger-Boos correction (Berger and Boos 1994), as in StatXact (Cytel 2007, Section 17.3.10).

`RRtosst()` gives intervals obtained by inverting two one-sided score tests; `RRotsst()` gives intervals obtained by inverting one two-sided score test. `RRtosst()` is thus more conservative, preserving at least $\alpha/2$ in each tail. These functions use a simple step search algorithm.

```
> RRotsst(c(4,24,12,28))

PF
95% interval estimates

    PF     LL     UL
0.6111 0.0148 0.8519
```

```
> RRtosst(c(4,24,12,28))

PF
95% interval estimates

   PF    LL    UL
0.611 0.012 0.902
```

# 3 Stratified designs

Methods for estimating a common *RR* from stratified or clustered designs depend on homogeneity with respect to the common parameter.

## 3.1 Gart's score method

Gart (1985) derived a score statistic for a common estimate of *RR* from designs with multiple independent strata, and he showed that it is identical to one proposed by Radhakrishnan (1965) from a different perspective.

`RRstr()` provides confidence intervals and a homogeneity test based on Gart's statistic.

Data may be input two ways, either using a formula and data frame, or as a matrix.

```
> RRstr(cbind(y,n) ~ tx + cluster(clus), Table6 , pf = F)

Test of homogeneity across clusters

stat         0.954
df          3
p          0.812

RR
95% interval estimates

            RR   LL   UL
starting  2.66 1.37 5.18
mle       2.65 1.39 5.03
skew corr 2.65 1.31 5.08
```

## 3.2 Mantel-Haenszel estimator

A widely-used heuristic method for sparse frequency tables is the weighted average approach of Mantel and Haenszel (1959).[1] MH interval estimates are based on the asymptotic normality of the log of the risk ratio. `RRmh()` utilizes the variance estimator given by Greenland and Robins (1985) for sparse strata. The resulting asymptotic estimator is consistent for both the case of sparse strata where the number of strata is assumed increasing, and the

[1] Kuritz et al. (1988) review the Mantel-Haenzel approach and point out its relationship to a method proposed by Cochran (1954), which was the basis of Rhadakrishnan's method (Radhakrishnan 1965), alluded to in section 3.1.

case of limited number of strata where the stratum size is assumed increasing. In the latter case, however, it is less efficient than maximum likelihood (Agresti and Hartzel 2000; Greenland and Robins 1985). Additional discussion may be found in Lachin (2000, Section 4.3.1), Landis et al. (2005), and Somes and O'Brien (2006). [2]

```
> RRmh(cbind(y,n) ~ tx + cluster(clus), Table6 , pf = F)
```

```
RR
95% interval estimates

  RR   LL   UL
2.67 1.37 5.23
```

## 3.3 Examples

For a full set of examples, see Chris Tong's vignette *Examples for Stratified Designs*.

# 4 Model based intervals

## 4.1 Logistic regression estimates

Intervals may be estimated from logistic regression models with `RRor()`. It takes the fit of a `glm()` object and estimates the intervals by the delta method.

```
> bird.fit <- glm(cbind(y,n-y) ~ tx - 1, binomial, bird)
> RRor(bird.fit)
```

```
95% t intervals on 4 df

PF
     PF      LL      UL
 0.5000 -0.0406  0.7598

      mu.hat    LL    UL
txcon  0.733 0.896 0.466
txvac  0.367 0.624 0.168
```

[2] SAS Proc FREQ provides MH interval estimates of *RR*. The other estimator calculated by Proc FREQ, which it calls "logit," is actually a weighted least squares estimator (Lachin 2000) that has a demonstrable and severe bias for sparse data (Greenland and Robins 1985). It should be avoided.

## 4.2 Estimating the dispersion parameter

The binomial GLM weights are

$$\frac{\hat{\pi}(1-\hat{\pi})}{a(\hat{\varphi})/n}$$

where $a(\hat{\varphi})$ is a function of the dispersion parameter.

### 4.2.1 Dispersion parameter $\varphi$

A simple estimator of the dispersion parameter, $\varphi$, may be estimated by the method of moments (Wedderburn 1974). It is given by `phiWt()`. This form of the dispersion parameter has $a(\varphi) = \varphi$, and $\varphi$ is estimated by $X^2/df$ , the Pearson statistic divided by the degrees of freedom.

Note that $\varphi$ is the same estimator as may be obtained by the `quasibinomial` family in `glm()` which is, in fact, what is used by `phiWt()` to reweight the original fit:

```
> phiWt(bird.fit, fit.only = F)$phi

     all
2.471592

> summary(update(bird.fit, family = quasibinomial))$disp

[1] 2.471592
```

`phiWt()` makes it easy to estimate *PF* intervals with a single command.

```
> # model weighted by phi hat
> RRor(phiWt(bird.fit))

95% t intervals on 4 df

PF
    PF     LL     UL
 0.500 -0.583  0.842

      mu.hat    LL     UL
txcon  0.733 0.943 0.3121
txvac  0.367 0.752 0.0997
```

It also allows different estimates of $\hat{\varphi}$ for specified subsets of the data.

```
> # model with separate phi for vaccinates and controls
> RRor(phiWt(bird.fit, subset.factor = bird$tx))

95% t intervals on 4 df

PF
    PF     LL     UL
 0.500 -0.645  0.848

      mu.hat    LL     UL
txcon  0.733 0.938 0.3330
txvac  0.367 0.767 0.0925
```

If you want to subtract a degree of freedom for each additional parameter, you can do that by entering the degrees of freedom as an argument to `RRor()`.

```
> # subtract 2 degrees of freedom
> RRor(phiWt(bird.fit, subset.factor = bird$tx), degf = 2)

95% t intervals on 2 df

PF
    PF     LL     UL
 0.500 -2.164  0.921

      mu.hat    LL     UL
txcon  0.733 0.975 0.1635
txvac  0.367 0.895 0.0377
```

### 4.2.2 Dispersion parameter $\tau$

When overdispersion is due to intra-cluster correlation, it may make sense to estimate the dispersion as a function of the intra-cluster correlation parameter $\tau$. In other words, $a(\varphi_{ij}) = 1 + \tau_j(n_{ij} - 1)$. `tauWt()` does this using the Williams procedure (Williams 1982).

```
> # model weighted using tau hat
> RRor(tauWt(bird.fit, subset.factor = bird$tx))

95% t intervals on 4 df

PF
    PF     LL     UL
 0.500 -0.645  0.848
```

```
      mu.hat    LL     UL
txcon  0.733 0.938 0.3330
txvac  0.367 0.767 0.0925
```

In this example the `tauWt()` estimates are the same as the `phiWt()` estimates. That is because the cluster sizes are all the same. Let's see what happens if we modify the `bird` data set. The `birdm` data set has the same cluster fractions but differing cluster sizes.

```
> # different cluster sizes, same cluster fractions
> birdm.fit <- glm(cbind(y,n-y) ~ tx - 1, binomial, birdm)
> RRor(tauWt(birdm.fit, subset.factor = birdm$tx))
```

```
95% t intervals on 4 df

PF
    PF     LL     UL
 0.490 -0.605  0.838

      mu.hat    LL    UL
txcon  0.737 0.942 0.328
txvac  0.376 0.764 0.101
```

Note that increasing cluster size can make things worse when there is intra-cluster correlation.

Now let's compare the weights from `phiWt()` and `tauWt()` with unequal cluster sizes. In the output below, $w$ represents $1/a(\hat{\varphi})$ and $nw$ is $n/a(\hat{\varphi})$

```
> # Compare phi and tau weights
> #
> phi.wts <-phiWt(birdm.fit,fit.only = F, subset.factor = birdm$tx)$weights
> tau.wts <- tauWt(birdm.fit,fit.only = F, subset.factor = birdm$tx)$weights
> w <- cbind(w.phi=phi.wts,w.tau=tau.wts,nw.phi=phi.wts*birdm$n,
+                   nw.tau=tau.wts*birdm$n)
> print(cbind(birdm[,c(3,1,2)],round(w, 2)), row.names=F)
```

```
  tx  y  n w.phi w.tau nw.phi nw.tau
 vac  1 10  0.32  0.35   3.20   3.55
 vac  8 20  0.32  0.21   6.40   4.13
 vac  9 15  0.32  0.26   4.80   3.92
 con  8 16  0.21  0.27   3.39   4.33
```

```
 con  8 10  0.21  0.38   2.12   3.82
 con 27 30  0.21  0.16   6.36   4.84
```

Look at the last two rows. Note that the `nw.phi` are directly proportional to $n$ within treatment group, while the `nw.tau` are not. With intra-cluster correlation, increasing cluster size does not give a corresponding increase in information.

### 4.3 Rao-Scott weights

Rao and Scott (1992) give a method of weighting clustered binomial observations based on the variance inflation due to clustering. They relate their approach to the concepts of design effect and effective sample size familiar in survey sampling, and they illustrate its use in a variety of contexts. `rsbWt()` implements it in the same manner as `phiWt()` and `tauWt()`. For more general use, the function `rsb()` just returns the design effect estimates and the weights.

```
> # model weighted with Rao-Scott weights
> RRor(rsbWt(birdm.fit, subset.factor = birdm$tx))
```

```
95% t intervals on 4 df

PF
    PF     LL     UL
 0.479 -0.314  0.793

      mu.hat    LL    UL
txcon  0.768 0.960 0.311
txvac  0.400 0.717 0.149
```

```
> # just the design effect estimates
> rsb(birdm$y, birdm$n, birdm$tx)$d
```

```
     con      vac
5.137107 2.500000
```

## 5 Likelihood based intervals

The `RRlsi()` function estimates likelihood support intervals for $RR$ by the profile likelihood (Royall 1997, Section 7.6).

Likelihood support intervals are usually formed based on the desired likelihood ratio, $1/k$ , often $1/8$ or $1/32$ . Under some conditions the log likelihood ratio may follow the chi-square distribution. If so, then $\alpha = 1 - F_{\chi^2}(2\log(k), 1)$. `RRsc()` will make the conversion from $\alpha$ to $k$ with the argument `use.alpha = T`.

```
> RRlsi(c(4,24,12,28))

1/8 likelihood support interval for PF

corresponds to 95.858% confidence
  (under certain assumptions)

PF
    PF     LL     UL
0.6111 0.0168 0.8859

> RRlsi(c(4,24,12,28), use.alpha = T)

1/6.826 likelihood support interval for PF

corresponds to 95% confidence
  (under certain assumptions)

PF
    PF     LL     UL
0.6111 0.0495 0.8792
```

# 6 Incidence ratio

The incidence is the number of cases per subject-time. Its distribution is assumed Poisson. Under certain designs, the incidence ratio (*IR*) is used as a measure of treatment effect. Correspondingly, $PF_{IR} = 1 - IR$ would be used as a measure of effect for an intervention that is preventive, such as vaccination. *IR* is also called incidence density ratio (*IDR*), and that is the term used in the following functions.

## 6.1 Score method

`IDRsc()` estimates a confidence interval for the incidence density ratio using Siev's formula (Siev 1994, Appendix 1) based on the Poisson score statistic.[3]

$$IDR = \widehat{IDR}\left\{1 + \left(\frac{1}{y_1} + \frac{1}{y_2}\right)\frac{z^2_{\alpha/2}}{2} \quad \pm \quad \frac{z^2_{\alpha/2}}{2y_1y_2}\sqrt{y_\bullet\left(y_\bullet z^2_{\alpha/2} + 4y_1y_2\right)}\right\}$$

```
> IDRsc(c(26,204,10,205), pf = F)

IDR
95% interval estimates

 IDR   LL   UL
2.61 1.28 5.34
```

## 6.2 Likelihood method

A likelihood support interval for *IDR* may be estimated based on orthogonal factoring of the reparameterized likelihood. (Royall 1997, Section 7.2) `IDRlsi()` implements this method.

Likelihood support intervals are usually formed based on the desired likelihood ratio, $1/k$ , often $1/8$ or $1/32$ . Under some conditions the log likelihood ratio may follow the chi square distribution. If so, then $\alpha = 1 - F_{\chi^2}(2\log(k), 1)$. `IDRlsi()` will make the conversion from $\alpha$ tp $k$ with the argument `use.alpha = T`.

```
> IDRlsi(c(26,204,10,205), pf = F)

1/8 likelihood support interval for IDR

corresponds to 95.858% confidence
  (under certain assumptions)

IDR
 IDR   LL   UL
2.61 1.26 5.88

> IDRlsi(c(26,204,10,205), pf = F, use.alpha = T)
```

[3] This formula was published in a Japanese journal (Sato 1988) several years before Siev. See also Graham et al. (2003) and Siev (2004).

```
1/6.826 likelihood support interval for IDR

corresponds to 95% confidence
  (under certain assumptions)

IDR
 IDR   LL   UL
2.61 1.30 5.69
```

# References

Agresti, A. and Hartzel, J. (2000), "Strategies for comparing treatments on a binary response with multi-centre data," *Statistics in Medicine*, 19, 1115–1139.

Berger, R. L. and Boos, D. D. (1994), "P values maximized over a confidence set for the nuisance parameter," *Journal of the American Statistical Association*, 89, 214–220.

Cochran, W. G. (1954), "Some methods for strengthening the common chi-square tests," *Biometrics*, 10, 417–451.

Cytel (2007), *StatXact 8 Procs User Manual*, Cytel Inc., Cambridge, MA.

Gart, J. J. (1985), "Approximate tests and interval estimation of the common relative risk in the combination of $2 \times 2$ tables," *Biometrika*, 72, 673–677.

Gart, J. J. and Nam, J. (1988), "Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness," *Biometrics*, 44, 323–338.

Graham, P. L., Mengersen, K., and Morton, A. P. (2003), "Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method," *Statistics in Medicine*, 22, 2071–2083.

Greenland, S. and Robins, J. M. (1985), "Estimation of a common effect parameter from sparse follow-up data," *Biometrics*, 41, 55–68, errata, 45:1323–1324.

Koopman, P. A. R. (1984), "Confidence intervals for the ratio of two binomial proportions," *Biometrics*, 40, 513–517.

Kuritz, S. J., Landis, J. R., and Koch, G. G. (1988), "A general overview of Mantel-Haenszel methods: applications and recent developments," *Annual Review of Public Health*, 9, 123–160.

Lachin, J. M. (2000), *Biostatistical Methods: The Assessment of Relative Risks*, New York: Wiley.

Landis, J. R., Sharp, T. J., Kuritz, S. J., and Koch, G. G. (2005), "Mantel-Haenszel methods," in *Encyclopedia of Biostatistics*, eds. Armitage, P. and Colton, T., Chichester: Wiley, vol. 4, 2nd ed., pp. 2937–2950.

Mantel, N. and Haenszel, W. (1959), "Statistical aspects of the analysis of data from retrospective studies of disease," *Journal of the National Cancer Institute*, 22, 719–748.

Miettinen, O. and Nurminen, M. (1985), "Comparative analysis of two rates," *Statistics in Medicine*, 4, 213–226.

Radhakrishnan, S. (1965), "Combination of results from several $2 \times 2$ contingency tables," *Biometrics*, 21, 86–98.

Ralston, M. L. and Jennrich, R. I. (1978), "DUD, A derivative-free algorithm for nonlinear least squares," *Technometrics*, 20, 7–14.

Rao, J. N. K. and Scott, A. J. (1992), "A simple method for the analysis of clustered binary data," *Biometrics*, 48, 577–585.

Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*, Boca Raton: Chapman and Hall.

Sato, T. (1988), "Confidence intervals for effect parameters from cohort studies based on efficient scores," *Japanese Journal of Applied Statistics*, 17, 43–54, in Japanese.

Siev, D. (1994), "Estimating vaccine efficacy in prospective studies," *Preventive Veterinary Medicine*, 20, 279–296.

— (2004), "Letter to the editor," *Statistics in Medicine*, 23, 693, (Note − Typographical error in formula: replace the two final minus signs with subscript dots).

Somes, G. W. and O'Brien, K. F. (2006), "Mantel-Haenszel statistics," in *Encyclopedia of Statistical Sciences*, eds. Kotz, S., Balakrishnan, N., Read, C. B., and Vidakovic, B., Hoboken: Wiley, vol. 7, 2nd ed., pp. 4483–4486.

Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method," *Biometrika*, 61, 439–447.

Williams, D. A. (1982), "Extra-binomial variation in logistic linear models," *Applied Statistics*, 31, 144–148.