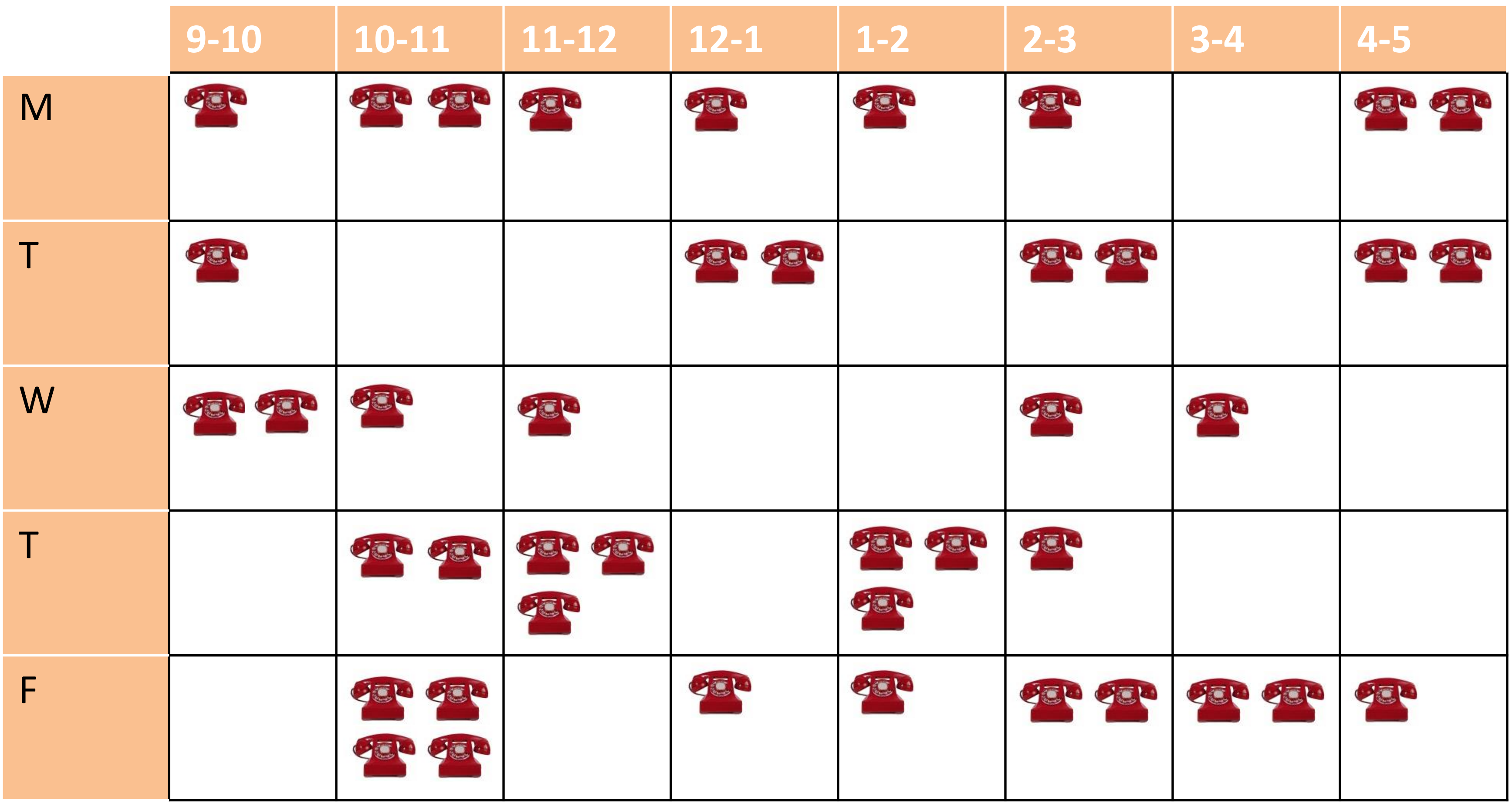
A dig into Poisson regression

1. # phone calls per hour



1. # phone calls per hour

- Can this be modelled using a Poisson distribution?
- Main assumption: mean = variance

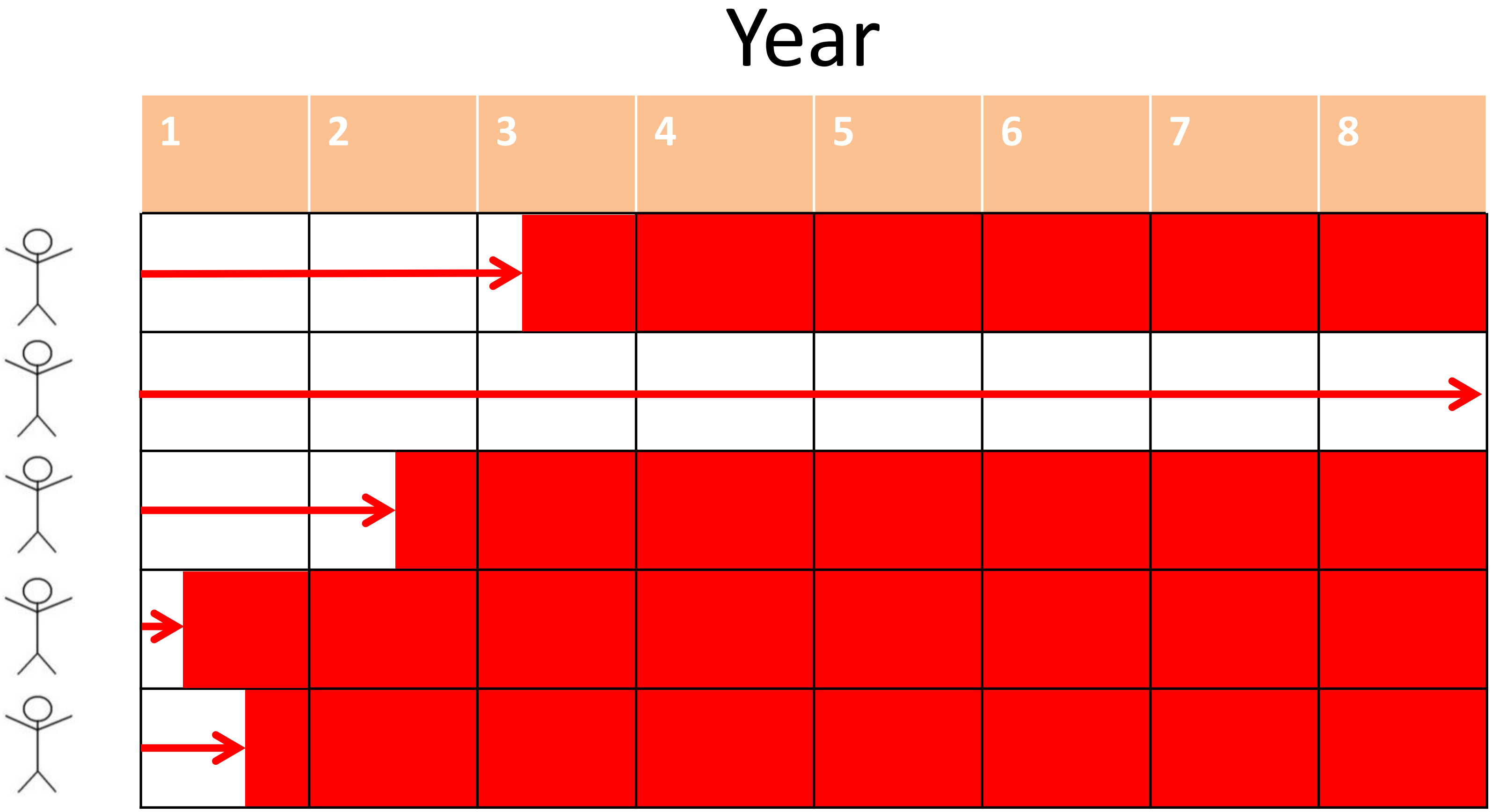
- Mean = $\frac{Y_1 + \cdots + YN}{N} = \frac{0+2+1+\cdots+}{40} = 1.05$

Units: hours

- Variance $= \frac{(Y_1-\mu)^2+\cdots+(Y_N-\mu)^2}{N}$

$$= \frac{(0-1.05\ )^2+(2-1.05\ )^2+\cdots+(\ 1-1.05\ )^2}{40} = 0.87$$

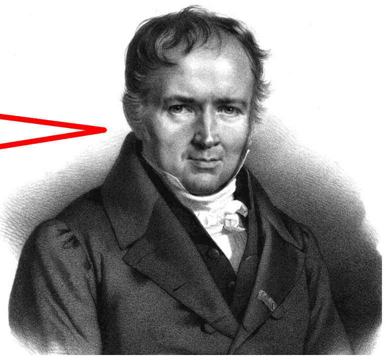
2. Incidence of Diabetes



2. Incidence of Diabetes

- Can this be modelled using a Poisson distribution?

- Main assumption: mean = variance



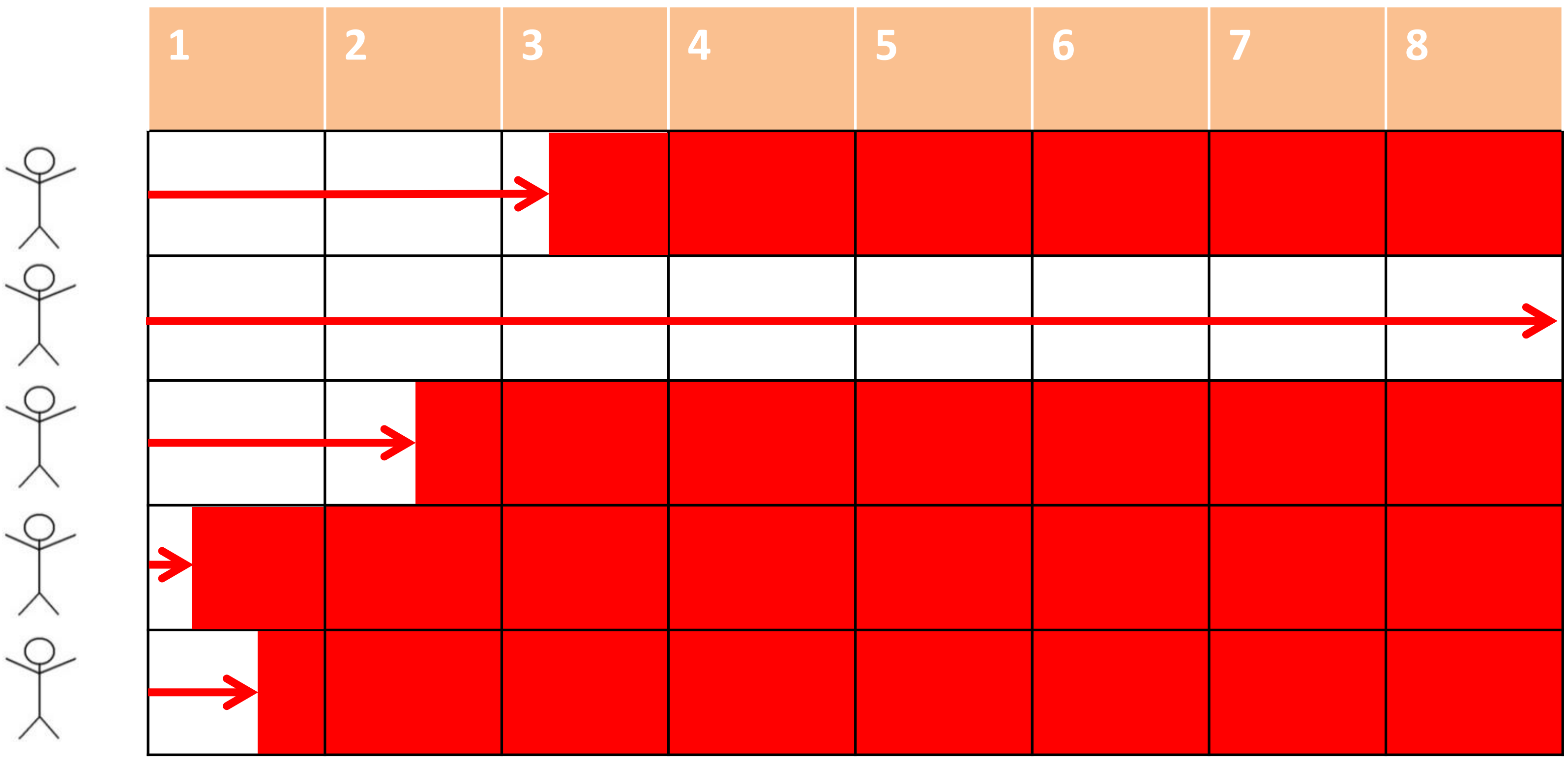
- Mean = $\frac{1+0+1+\cdots+1}{14.18} = 0.28$

Units: Person-years

- We have a problem calculating the variance, as some of our units are not whole.

2. Incidence of Diabetes

Year



Variance = $\dfrac{(Y_1-\mu)^2+\cdots+(Y_N-\mu)^2}{N}$

2. Incidence of Diabetes

- Let Y be the number of incident diabetes diagnoses per person-year.

- Suppose we want to test whether Y ~ Po(0.28)

- Let:
  - Y_i be the number of incident diabetes diagnoses in person i (0/1);
  - T_i be the number of person years followed up.

Then we’re testing whether $Y_i \sim Po(0.28 * T_i)$

2. Incidence of Diabetes

- Each person having a different Poisson distribution makes it hard to test everyone at the same time.
- Somehow need to standardise each person’s contribution.
- One way of doing this is to **divide by the variance**:

2. Incidence of Diabetes

- Going back to our original criteria:

$$\left\{\frac{(Y_1 - E[Y_1])^2 + \cdots + (Y_N - E[YN])^2}{N}\right\} = \text{var}[Yi]$$

$$\left\{\frac{(Y_1 - E[Y_1])^2 + \cdots + (Y_N - E[YN])^2}{\text{var}[Yi]}\right\} = N$$

$$\left\{\frac{(Y_1 - E[Y_1])^2}{\text{var}[Yi]} + \cdots + \frac{(Y_N - E[YN])^2}{\text{var}[Yi]}\right\} = N$$

- If the var[Yi] are now different, our criteria becomes:

$$\left\{\frac{(Y_1 - E[Y_1])^2}{\text{var}[Y_1]} + \cdots + \frac{(Y_N - E[YN])^2}{\text{var}[YN]}\right\} = N$$

GLM Residuals

- The square-root of each of the individual components of the equation below are **Pearson residuals**, p_i, with the left-hand side, $\sum(p_i{}^2)$, often reported as the **Pearson goodness-of-fit (GoF)** statistic.

- Turns out the equation below is only valid if $E[Y_i] > 5$, which is unfortunate for epi studies where each person only has max value of 1.

- Alternative residuals for GLMs are **deviance residuals**, d_i, based on likelihood ratios:

$$d_i = \pm\sqrt{-2li}$$

where l_i is the contribution of Y_i to the log-likelihood

- The **sum of deviance residuals squared**, $\sum(d_i{}^2)$, is more commonly compared to N to test for model fit.

Stata example

- Simulate: 100 men ~ Po(10)
  100 women ~ Po(5)

- Outcome: diabetes

- Followed up for 10 years or until they get diabetes

Stata example

```
. poisson diabetes, e(fu) irr

Iteration 0:   log likelihood = -312.24417
Iteration 1:   log likelihood = -312.24417

Poisson regression                                Number of obs   =        200
                                                  LR chi2(0)      =      -0.00
                                                  Prob > chi2     =          .
Log likelihood = -312.24417                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
    diabetes |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .3153936   .0227025   -16.03   0.000     .2738937    .3631815
      ln(fu) |          1  (exposure)
------------------------------------------------------------------------------

. estat gof

         Deviance goodness-of-fit =  238.4883
         Prob > chi2(199)         =    0.0291

         Pearson goodness-of-fit  =  1224.745
         Prob > chi2(199)         =    0.0000
```

Overdispersed

Stata example 2

```
. poisson diabetes sex, e(fu) irr

Iteration 0:   log likelihood = -300.82152
Iteration 1:   log likelihood = -300.82028
Iteration 2:   log likelihood = -300.82028

Poisson regression                          Number of obs   =        200
                                            LR chi2(1)      =      22.85
                                            Prob > chi2     =     0.0000
Log likelihood = -300.82028                 Pseudo R2       =     0.0366

------------------------------------------------------------------------------
    diabetes |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .5003176    .0720747    -4.81   0.000     .3772449    .6635415
       _cons |   .9337626    .2103614    -0.30   0.761     .6004471    1.452106
      ln(fu) |          1  (exposure)
------------------------------------------------------------------------------

. estat gof

         Deviance goodness-of-fit =  215.6406
         Prob > chi2(198)         =    0.1855

         Pearson goodness-of-fit  =  925.3769
         Prob > chi2(198)         =    0.0000
```

Negative Binomial

- Is basically a Poisson model with an extra dispersion parameter, α.
- Rather than assuming the variance is μ, it assumes the **variance is** $\mathbf{(1+\alpha)}\boldsymbol{\mu}$.
- In stata, negative binomial regression output will test whether α is significantly different to zero.

Negative Binomial

```
. nbreg diabetes, e(fu) irr

Negative binomial regression                      Number of obs   =        200
                                                  LR chi2(0)      =      -0.00
Dispersion      = mean                            Prob > chi2     =          .
Log likelihood = -323.76582                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
    diabetes |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   7.815708   .8486891    18.94   0.000       6.3174    9.669372
      ln(fu) |          1  (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |  -1.890833   .5505114                     -2.969816   -.8118507
-------------+----------------------------------------------------------------
       alpha |    .150946   .0830975                      .0513128    .4440355
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =     4.55 Prob>=chibar2 = 0.016
```

Negative Binomial

```
. nbreg diabetes sex, e(fu) irr

Negative binomial regression                    Number of obs     =        200
                                                LR chi2(1)        =      22.85
Dispersion        = mean                        Prob > chi2       =     0.0000
Log likelihood = -300.8202                      Pseudo R2         =     0.0366

------------------------------------------------------------------------------
    diabetes |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .5003174   .0720741    -4.81   0.000     .3772457    .6635398
       _cons |   .9337779   .2103631    -0.30   0.761      .600459    1.452124
      ln(fu) |          1  (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |  -18.86967    509.578                     -1017.624    979.8849
-------------+----------------------------------------------------------------
       alpha |   6.38e-09   3.25e-06                             0           .
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 1.7e-04 Prob>=chibar2 = 0.495
```

More extreme example

- Simulate: 100 men ~ Po(**100**)

  100 women ~ Po(5)

- Outcome: diabetes

- Followed up for 10 years or until they get diabetes

More extreme example

```
. poisson diabetes, e(fu) irr

Iteration 0:   log likelihood = -473.93622
Iteration 1:   log likelihood = -473.93622

Poisson regression                              Number of obs     =        200
                                                LR chi2(0)        =       0.00
                                                Prob > chi2       =          .
Log likelihood = -473.93622                     Pseudo R2         =     0.0000

------------------------------------------------------------------------------
    diabetes |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |    10.1021    .7143265    32.71   0.000     8.794735    11.60382
       ln(fu) |          1  (exposure)
------------------------------------------------------------------------------

. estat gof

         Deviance goodness-of-fit =  547.8724
         Prob > chi2(199)         =    0.0000

         Pearson goodness-of-fit  =  4721.387
         Prob > chi2(199)         =    0.0000
```

More extreme example

```
. nbreg diabetes, e(fu) irr

Negative binomial regression                    Number of obs   =        200
                                                LR chi2(0)      =       0.00
Dispersion     = mean                           Prob > chi2     =          .
Log likelihood = -421.48596                     Pseudo R2       =     0.0000
```

diabetes	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	42.65677	7.250667	22.08	0.000	30.57056	59.52133
ln(fu)	1	(exposure)				
/lnalpha	.1639804	.1490359			-.1281246	.4560854
alpha	1.178191	.1755928			.8797437	1.577885

```
Likelihood-ratio test of alpha=0:  chibar2(01) =  104.90 Prob>=chibar2 = 0.000
```

More extreme example

```
. poisson diabetes sex, e(fu) irr

Iteration 0:   log likelihood = -346.36248
Iteration 1:   log likelihood = -346.35994
Iteration 2:   log likelihood = -346.35994

Poisson regression                                Number of obs   =        200
                                                  LR chi2(1)      =     255.15
                                                  Prob > chi2     =     0.0000
Log likelihood = -346.35994                       Pseudo R2       =     0.2692

------------------------------------------------------------------------------
    diabetes |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   12.24406   1.731571    17.71   0.000     9.279985    16.15487
       _cons |   .4462232   .0997785    -3.61   0.000     .2878841    .6916504
      ln(fu) |          1  (exposure)
------------------------------------------------------------------------------

. estat gof

         Deviance goodness-of-fit =  292.7199
         Prob > chi2(198)         =    0.0000

         Pearson goodness-of-fit  =  2447.731
         Prob > chi2(198)         =    0.0000
```

More extreme example

```
. nbreg diabetes sex, e(fu) irr

Negative binomial regression                     Number of obs   =        200
                                                  LR chi2(1)      =     179.88
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -331.54513                       Pseudo R2       =     0.2134
```

diabetes	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	15.6766	2.644231	16.32	0.000	11.26358	21.81862
_cons	.4127013	.1101791	-3.32	0.001	.2445626	.6964367
ln(fu)	1	(exposure)				
/lnalpha	-1.520862	.3265859			-2.160959	-.8807654
alpha	.2185234	.0713667			.1152146	.4144656

Likelihood-ratio test of alpha=0: chibar2(01) = 29.63 Prob>=chibar2 = 0.000

2. Incidence of Diabetes

We divide each person's variance-contribution by the variance of their hypothesised distribution, add them all up, and see if it's roughly equal to N.

$$\left\{\frac{(Y_1 - E[Y_1])^2}{\mathrm{var}[Y_1]} + \cdots + \frac{(Y_N - E[YN])^2}{\mathrm{var}[YN]}\right\} = N$$