The calibration software CALMAR – What is it?

Kevin McCormack
Central Statistics Office Ireland

Contents

1. Introduction

Calibration is an approach often used by survey practitioners to improve estimates from sample surveys when auxiliary information about the population that is under study is available. The key feature is the modification of the sample survey weights to reproduce from the sample population characteristics, namely population totals and category frequencies. For example, in a sample survey of a human population age and sex are natural ancillary variables. The distribution of the human population by age and sex is often known from other statistical sources such as a census or a population register and by proper modification of the survey weights, the population structure may be exactly reproduced by the sample. For variables in the survey correlated with the ancillary information, higher precision in estimates is usually obtained on application of the new calibrated weights.

Calibration has its origin in the raking procedures introduced by Deming and Stephan (1940). In later years, the calibration technique has received considerable attention from official statistical authorities. The interest for the approach has grown since Deville and Särndal (1992) showed the asymptotic equivalence of calibration to the generalised regression estimator (Cassel, Särndal and Wretman, 1976), thereby providing a way to establish the statistical properties of calibrated estimators.

In the following, a short presentation concerning the calibration software CALMAR, including the theoretical principles underlying the calibration methods the software implements, is given. Short discussions on the mechanisms for non-response and contingency tables are also contained in the annexes.

2. The calibration technique

2.1. Formulation and solution to the problem

For a long time, auxiliary information has been used at the estimation stage in order to obtain estimates of higher precision. For example, the generalised regression estimator (GREG) uses auxiliary variables with known population totals coming from exogenous sources. It is well-known (Särndal, Swensson and Wretman, 1992) that the more those variables are correlated with the study variable, the better the precision of the GREG estimator is.

Deville and Särndal (1992) introduced a single harmonised framework for the use of auxiliary data at the estimation stage. Let us consider a finite population U, or universe, of N distinct units. We can denote the variable under study, say males, by y. Now suppose that we want to estimate the total Y of males summing over the entire population from a sample s of size n drawn from U according to a given sampling design. To estimate the population total in such circumstances, a linear estimator is used:

$$\hat{Y}_d = \sum_{k \in s} d_k \cdot y_k$$

Where d_k is the *sampling weight* of k.

Suppose now that there exist J auxiliary variables $x_1 \ldots x_j \ldots x_J$, called *calibration variables*, with known population totals (in the case of numerical variables) or marginal counts (in the case of categorical variables). Without loss of generality, we can assume that all the calibration variables are numerical (otherwise, we consider the 0/1 variables for each category).

We seek new sampling weights ω_k that are "as close as possible" (as determined by a certain distance function) to the initial weights d_k. These ω_k are calibrated on the totals X_j of the J auxiliary variables; in other words they verify the calibration equations:

$$\boxed{\forall\, j = 1 \ldots J \qquad \sum_{k \in s} \omega_k \cdot x_{jk} \;=\; X_j}. \qquad (1)$$

The solution to this problem is given by: $\omega_k = d_k \cdot F\left(x_k' \cdot \lambda\right)$

Where:

- $x_k' = (x_{1k} \ldots x_{jk} \ldots x_{Jk})$
- λ is the vector of the J Langrage multipliers associated with constraint (1) above
- F is a function – *the calibration function* – whose terms depend on the distance function that is used.

Vector λ is determined by the solution to the non-linear system of J equations in J unknowns resulting from the calibration equations:

$$\sum_{k \in s} d_k \cdot F\left(x_k' \lambda\right) x_k \quad = \quad X$$

Where: $X' = \left(X_1 \ldots X_J\right)$

Finally, the *calibrated estimator* of the total Y for the study variable y is:

$$\hat{Y}_{CAL} \;=\; \sum_{k \in s} \omega_k \cdot y_k$$

For example, if we take *one known* total X, we then calibrate by constructing weights ω_i such that:

$$\sum_{k \in s} \omega_k \cdot x_k \quad = \quad X$$

Where the new weights ω_k are as close as possible to the old weights d_k.

This can be done by minimising the following Chi-squared quadratic distance:

$$\sum_{i \in s} \frac{(\omega_k - d_k)^2}{d_k}$$

the solution of which gives the new weights:

$$\omega_k = d_k \cdot \left[1 + \frac{\left(X - \hat{X}_d \right)}{\hat{X}_s} x_k \right]$$

Where $\hat{X}_s = \sum_{k \in s} d_k x_k^2$

2.2. Statistical properties of calibrated estimators

Deville and Särndal (1992) showed that under general assumptions a calibrated estimator is asymptotically equivalent to the GREG estimator (where the calibration variables are the "regression" variables), in the sense that:

$$N^{-1} \cdot (\hat{Y}_{CAL} - \hat{Y}_{GREG}) = O_P\left(\frac{1}{n}\right)$$

Thus, a calibrated estimator has a bias of order 1/*n*, which is asymptotically negligible. Besides, its variance can be easily estimated by substituting the study variable y with the regression residuals. In particular, if the regression model is good (i.e. the calibration variables are correlated with the study variable), the variance can be substantially decreased, at the expense of a very small bias.

2.3. Calibration and non-response

Although it has been introduced originally as a variance reduction method, the calibration can be used for both increasing the precision of estimates and reducing non-response bias. Contrary to the classical approach for dealing with non-response, which consists of making homogeneous response groups[1], the auxiliary variables have to be known for the responding units only. However, their population totals must be known as well, which is quite restricting.

To solve this problem, a "super-generalised" calibration theory was developed (Deville, 2000). The idea is to calculate weights of the form:

$$\omega_k = d_k \cdot F(x_k' \cdot \lambda)$$

[1] The sample is divided into cells where the response propensity is assumed to be uniform. Then, the response rate within a cell is an estimate (maximum of likelihood) for that propensity.

Where $z_k = (z_{1k}\ldots z_{Jk})'$ is a vector of non-response variables. These weights must satisfy calibration equations based on calibration variables $x_1\ldots x_J$

$$\sum_{k\in s} \omega_k \cdot x_k \quad = \quad X$$

Thus, there is no need to know population totals of non-response variables. In particular, survey variables can be directly used for correcting non-response. This is interesting for social surveys (EU-SILC, LFS…) where non-response is expected to be correlated with the survey target variables.

This option is offered in CALMAR2.

2.4. Choice of the calibration information

According to 2.2., the better the calibration variables fit the survey variables, the better the gain in precision is. Nevertheless, the choice of the information to be used in calibration is not that easy. It has been implicitly assumed in the previous sections that both the calibration variables and the calibration totals are exact, i.e. error-free. Of course, in practice, this assumption does not hold. Errors in auxiliary information may severely damage calibrated weights, as shown in the two next examples.

Example 1: Suppose that the values of a calibration variable, which are collected during the survey, are systematically under-estimated. That might happen for instance if the measurement tool that is used is defective or does not fit. On the other hand, the corresponding calibration total, which comes from an exogenous source, is assumed to be exact. It is clear that the consequence of using such a variable in calibration is the calibrated weights will be biased (over-estimated).

Example 2: Suppose now that we calibrate to frequency counts by activity status. The survey collects activity status pertaining to the survey year. On the other hand, the frequency counts which are available pertain to a previous year and were set up on the basis of another activity status classification. The consequence of using "inconsistent" information on activity status is it makes calibrated weights biased as well.

In short, we can say calibration is a quite powerful technique for improving the statistical properties of estimators, provided *caution* is taken as regards the information that is used for.

Another practical case is when the calibration totals come from another survey. For instance, the Irish Central Statistics Office used estimates from the LFS in order to calibrate the EU-SILC data. Using estimates as calibration totals instead of the "exact" values is generally harmless, provided:

- Those estimates can be regarded as (almost) unbiased and come from a sample which has at least the same size. If that size is bigger, one might expect a better precision in estimates.
- Both surveys measure the same information. In other words, the data collected in both surveys are "consistent" (see example 2 above).

3. CALMAR

CALMAR[2] is a SAS macro program that implements the calibration methods developed by Deville and Särndal (1992). The program calculates new sampling weights, based on the method described in 2.1. CALMAR offers four calibration methods, corresponding to four different distance functions.

3.1. The calibration methods

a. The linear method

It is based on the Chi-squared quadratic distance:

$$\sum_{i \in S} \frac{(\omega_k - d_k)^2}{d_k}$$

The linear method is always convergent and the convergence is fast. However, the calibrated weights can take negative values, which is not desirable. Moreover, the distribution of the calibrated weights is not bounded and can be highly skewed.

b. The exponential method

It is based on the following distance:

$$\omega_k \cdot log\left(\frac{\omega_k}{d_k}\right) - \omega_k + d_k$$

Contrary to the previous one, this method always leads to positive weights. However, they are still not bounded and can take quite large values.

c. The logit and the truncated linear method

Both of these methods are "bounded", which means they provide lower and upper limits (*LO* and *UP)* on the weight ratios w_k / d_k , usually referred to as the *g-weights*. Thus, they allow controlling the range of the distribution of weight ratios.

One has to keep in mind that the choice of the calibration limits *LO* and *UP* is not free and depends on the calibration variables which are chosen: the limits must be adjusted taking into account the differences between the estimates based on the old weights and the benchmark totals that the new weights are to reproduce, so CALMAR can find a solution within the constraints applied to the problem.

In practice, those limits are determined by "guess and check": we start with a small interval [*LO*,*UP*] and we enlarge it until CALMAR finds a solution.

[2] CALMAR is the acronym of CALibration on MARgins.

3.2. Choice of the calibration method

When using CALMAR, one has to choose one of the four methods that CALMAR offers. The choice of the method is generally based on *empirical criteria*. Indeed, according to 2.2, calibrated estimators are asymptotically equivalent to the GREG estimator so have same bias and variance. However, although all calibration methods (i.e. all distance functions) are *on average* equivalent, they do not produce the same results for a given sample.

For example, the linear method can produce negative or abnormally large weights. Estimates based on skewed weight distributions may take unexpected values, especially those for small domains. Thus, a caution rule is applied in practice: the "best" method is the one to which estimates are the least sensitive. In this regard, the "bounded" methods (logit and truncated linear) are generally preferred as they prevent from extreme weights. However, it is worth testing all the methods in order to see their impact on the weight distribution.

4. Examples

As a macro SAS program, CALMAR is easy to handle. The software prints summary statistics on the calibrated weights and on the weight ratios (g-weights) to help the user to assess the quality of the procedure. It is essential to look at those outputs systematically after running CALMAR.

Two preliminary datasets are needed to make the program run:

- *The "sample" dataset* (parameter DATAMEN), with the values taken by the calibration variables on each sample unit.

- *The "benchmark" dataset* (parameter MARMEN), with the totals and marginal counts for the variables.

***Example 1*: one numerical variable**

We consider a population of size N=120 from which a simple random sample of size n=15 is drawn. We also consider a numerical variable *x*. The values taken by that variable on the sampled units (we assume no non-response) are: 1, 1, 1, 2, 2, 4, 6, 7, 8, 11, 12, 10, 15, 20 and 50. The weighted[3] sum of *x* is $150 \times 8 = 1200$.

Suppose now that the exact population total of that variable is known and equals 1400. We would like to use CALMAR to calibrate the sample to this auxiliary total. In other words, we would like to slightly modify the initial weights so the sum of *x* based on the new weights is 1400.

The SAS code for creating the preliminary datasets and running CALMAR is given in the next page.

[3] The weights are the "design" weights N / n = 120 / 15 = 8

```
/*****************************/
/* Library containing CALMAR */
/*****************************/

libname calm 'D:\osiergu\General Files\CALMAR2_V9';
options mstored sasmstore=calm;


/*********************************/
/* Creation of the sample dataset */
/*********************************/

data sample;
 input ident $ x weight;
 cards;
a 1 8
b 1 8
c 1 8
d 2 8
e 2 8
f 4 8
g 6 8
h 7 8
i 8 8
j 11 8
k 12 8
l 10 8
m 15 8
n 20 8
o 50 8
;


/************************************/
/* Creation of the benchmark dataset */
/************************************/

data benchmark;
 input var $ n mar1;
 cards;
x 0 1400
;


/******************/
/* Call to CALMAR */
/******************/

%CALMAR2(
DATAMEN=sample,
POIDS=weight,
IDENT=ident,
MARMEN=benchmark,
M=1,
DATAPOI=tab_wght,
POIDSFIN=cal_weights
);


proc print data=tab_wght noobs;
```

```
run;
```

The following dataset, with the resultant weights, is printed:

ident	cal_ weights
a	8.0546
b	8.0546
c	8.0546
d	8.1091
e	8.1091
f	8.2182
g	8.3273
h	8.3819
i	8.4364
j	8.6001
k	8.6547
l	8.5456
m	8.8183
n	9.0911
o	10.7278

CALMAR increased the design weights so the benchmark total of 1400 can be reached. The design weight of the unit *o* was most increased as the unit carries the highest value for the variable *x*. Conversely, the design weights of *a*, *b* and *c* remained almost the same as the values taken by *x* on these units are small.

The solution worked out by the software is clearly one possible solution within the constraints of the problem, but there exist an infinite number of solutions (provided the calibration equations are consistent).

The macro parameter POIDS contains the weighting variable. The macro parameter IDENT contains the name of an identifying variable for the units in the sample dataset. Although it is an optional one, it should always be filled in. The macro parameter M is the number of the calibration method that is used (1: linear, 2: raking ratio, 3: logit and 4: truncated linear).

CALMAR will create a new dataset (parameter DATAPOI) which contains the calibrated weights (parameter POIDSFIN).

Example 2: one numerical variable and one categorical variable

We consider now the same variable *x* as in the first example plus a categorical variable *y* with three modalities, namely *a*, *b* and *c*. The values taken by that variable are: *a*, *a*, *b*, *b*, *c*, *a*, *c*, *c*, *b*, *a*, *a*, *b*, *c*, *c* and *b*. The estimated marginal counts for *a*, *b* and *c*, respectively, are: 40, 40 and 40.

Suppose now that the exact values for the marginal counts are available: 38, 41 and 39. We would like to run CALMAR to calibrate the sample to both the population total of *x* (i.e. 1400) and the marginal counts of *y* (38, 41 and 39).

```
/*****************************/
/* Library containing CALMAR */
/*****************************/

libname calm 'D:\osiergu\General Files\CALMAR2_V9';
options mstored sasmstore=calm;


/*********************************/
/* Creation of the sample dataset */
/*********************************/

data sample;
 input ident $ x y $ weight;
 cards;
a 1 a 8
b 1 a 8
c 1 b 8
d 2 b 8
e 2 c 8
f 4 a 8
g 6 c 8
h 7 c 8
i 8 b 8
j 11 a 8
k 12 a 8
l 10 b 8
m 15 c 8
n 20 c 8
o 50 b 8
;


/************************************/
/* Creation of the benchmark dataset */
/************************************/

data benchmark;
 input var $ n mar1 mar2 mar3;
 cards;
x 0 1400 . .
y 3 38 41 39
;


/******************/
/* Call to CALMAR */
/******************/

%CALMAR2(
DATAMEN=sample,
POIDS=weight,
IDENT=ident,
MARMEN=benchmark,
M=1,
DATAPOI=tab_wght,
POIDSFIN=cal_weights
```

```
);

proc print data=tab_wght noobs;
run;
```

We get the following dataset:

ident	cal_ weights
a	7.0996
b	7.0996
c	6.8240
d	6.9282
e	6.9661
f	7.4124
g	7.3830
h	7.4873
i	7.5537
j	8.1421
k	8.2463
l	7.7622
m	8.3212
n	8.8424
o	11.9319

Example 3: the Irish EU-SILC

The Irish Central Statistics Office applied CALMAR to produce calibrated weights for the EU-SILC survey. The variables in *Table 1* below were used in the production of the calibrated weights.

Table 1: Calibration variables – EU-SILC

Sex by Age

Males	*Females*
m1 = male 0-14	f1 = female 0-14
m2 = male 15-34	f2 = female 15-34
m3 = male 35-64	f3 = female 35-64
m4 = male 65+	f4 = female 65+

Region – NUTS 2	*Household Type*
1 = Border	1 = I adult, no children
2 = Midlands	2 = 2 adults, no children
3 = West	3 = 3+ adults, no children
4 = Dublin	4 = 1 adult with children
5 = Mid-East	5 = 2 adults, 1-3 children
6 = Mid-West	6 = Other households with children

7 = South-East
8 = South-West

The following plots shows the distribution of the weight ratios[4] for each of the 4 methods offered by CALMAR

Linear method (M=1)

```
                  Histogram                    #          Boxplot
 4.3+*                                         1             *
    .
    .
    .
    .*                                         2             *
    .*                                         5             0
    .*                                         3             0
    .*                                         4             0
    .*                                        20             0
    .**                                       35             0
    .***                                      52             |
    .****                                     84             |
    .************                            276             |
    .*****************************           725             |
    .*********************************************  1154    +-----+
    .*************************                 643    *--+--*
    .***************************               681    |     |
    .****************************************      1042    +-----+
    .************************                  607             |
    .*****                                    108             |
    .**                                        27             |
    .*                                          6             |
    .*                                          1             0
-0.3+*                                          1             0
     ----+----+----+----+----+----+----+----+----+--
     * may represent up to 25 counts
```

Exponential method (M=2)

```
                  Histogram                    #          Boxplot
6.75+*                                         1             *
    .
    .
    .
    .*                                         2             *
    .*                                         2             *
    .*                                         5             *
    .*                                         8             0
    .**                                       46             0
    .*****                                   166             |
    .*********************************       1456    +-----+
    .*********************************************  1941    *--+--*
    .******************************************    1845    +-----+
```

[4] Year: 2004

```
0.25+*                                                  5                |
     ----+----+----+----+----+----+----+----+----+---
     * may represent up to 41 counts
```

Logit method (M=3) (LO=0.4, UP=2.2)

```
                      Histogram                         #            Boxplot
2.15+*********************************                 382              |
    .*********                                         107              |
    .*****                                              50               |
    .************                                      137              |
    .*********************                             251              |
    .************************************              458            +-----+
    .*********************************                 414            |     |
    .************************************              444            |     |
    .***********************                           274            |     |
    .*******************************                   374            *--+--*
    .******************                                222            |     |
    .*********************************                 397            |     |
    .****************************************          504            |     |
    .***********************************************   554            +-----+
    .*****************************************         546              |
    .***********************                           272              |
    .*                                                  8              |
0.45+*******                                            83               |
     ----+----+----+----+----+----+----+----+----+--
     * may represent up to 12 counts
```

Truncated linear method (M=4) (LO=0.4, UP=2.2)

```
                      Histogram                         #            Boxplot
2.25+*******************************                   363              |
    .*****                                              58               |
    .**                                                 16               |
    .*******                                            84               |
    .***********                                       124              |
    .*****************                                 199              |
    .************************************              433              |
    .*****************************                     339            +-----+
    .****************************************          469            |     |
1.35+***********************************               426            |     |
    .*********************************                 405            *--+--*
    .*****************************                     349            |     |
    .************************************              449            |     |
    .***********************************               438            +-----+
    .***********************************************   542              |
    .*********************************                 393              |
    .******************                                228              |
    .******                                             71               |
0.45+********                                           91               |
     ----+----+----+----+----+----+----+----+----+-
     * may represent up to 12 counts
```

It is clear that the shapes of the plots vary depending on the calibration method. The linear method has produced some negative weights, which is not desirable from a practical point of view. On the contrary, the exponential method has produced only positive weights. However, some of the ratios take pretty high values (6.75).
As their corresponding plots show, the two "bounded" methods (logit and truncated linear) have made the weight ratios bounded between 0.4 and 2.2, which is more acceptable from a practical point of view. We also observe that the g-weights are concentrated next to the bounds, which is typical of that kind of methods.

In short, we can say that the "bounded" methods have distinct advantages, which make them interesting to survey practitioners. However, it is worth considering the exponential method too: even though it has lead to some outlying ratios, their share is quite small. This kind of distribution may sometimes appear to have more advantages than distributions where most of the weight ratios are concentrated around two fixed values.

The lesson is there is no definite answer *a priori* regarding the choice of the calibration method. It has to be based on empirical criteria: impact on the indicators, shape of the distribution of weight ratios, of final weights…

5. Conclusion

Deville and Särndal (1992) showed the asymptotic equivalence of calibration to the generalised regression estimator (Cassel, Särndal and Wretman, 1976), thereby providing a way to establish the properties of calibrated estimators.

There are some distinct advantages with calibration. Auxiliary information is incorporated in the weights. The obtained estimates are “consistent” with known information; usually this applies to totals. There is a correction for non-response, if non-response is related to the auxiliary information, and there is a reduction of variances for totals if variables are correlated.

Among the problems of the approach is that negative weights may appear. This is against our intuition. Modifications are possible at the expense of a more complicated procedure.

Annex I - The 2006 Irish EU-SILC

1. The survey

The 2006 Irish EU-SILC is a voluntary sample survey of private households. The Irish EU-SILC survey is designed to cover both cross-sectional and longitudinal components. This type of design has been titled “The integrated survey”. The basic unit of the survey will be the household. The cross-sectional component of the 2006 survey will have an effective sample size of some 6,000 households and the 2006 longitudinal component an effective sample size of some 3,500 households. The annual cross-sectional estimates will be produced from a rotational design, whereby 25% of the sample is rotated from one year to the next and retaining the other part unchanged. Each individual aged 16 years or over in a household must participate in the survey in order for a household to be accepted as a valid response.

The data collection element of the 2006 Irish EU-SILC started in January 2006 and will end in December 2006. The survey is continuous. The interviews are spread evenly over the year (i.e. on average 115 households surveyed each week). The survey will be conducted by face to face interviewing using laptops (BLAISE-CAPI)

2. Sample design

The sample is stratified for operational and economy reasons to distinguish between two different types of survey areas, namely: (a) town survey areas located in towns with 1,000 inhabitants or more (b) country survey areas covering towns with less than 1,000 inhabitants and all rural areas. A two-stage sample design is used. This comprises of a first stage sample of 2,600 blocks (or survey areas) randomly selected at county level to proportionately represent the following eight strata relating to population density

1. County Borough
2. Suburbs of County Boroughs
3. Environs of County Boroughs
4. Towns 10,000 +
5. Towns 5,000 - 10,000
6. Towns 1,000 – 5,000
7. Mixed Urban/Rural Areas
8. Rural Areas

Each block is selected to contain, on average, 75 households. The second sampling stage will involve the random selection of two independent samples of one original and two substitute households for each survey area. The original sample household constitutes the quota of co-operating households to be realised in each survey area and the interviewers systematically approach as many substitute households as is necessary to realise their quotas. In this fashion, variations in response by region and town size will be controlled.
The data collected as part of the fieldwork component of the EU-SILC is checked, verified and edited by CSO HQ staff.

3. Field Work

The Household Survey Collection Unit (HSCU) of the CSO has the responsibility for undertaking the fieldwork and data processing elements of the Irish 2006 EU-SILC programme. The HSCU consists of two teams, HQ and Field Staff. Field Staff members are trained by HQ staff in all aspects of household surveying, which includes interview techniques, laptop PC and time management. Modern distance management techniques are utilised by the HQ staff to supervise the work of the Field Staff.

The sample of households, and members therein, to be surveyed are provided to the Field Staff by HQ. It is the responsibility of the Field Staff to recruit the selected household in to the 2006 EU-SILC programme. The selected households are informed of their selection and the benefits of their participation by mail or phone (if a household had participated in the 2003, 2004 or 2005 EU-SILC programmes) prior to a Field Staff member calling to interview.

In the Irish EU-SILC programmes all members of the sample households are interviewed and the Field Staff have the responsibility to schedule the interviews to meet this requirement.

Reporting systems to indicate when to survey a particular household and whether or not such an interview is completed are features of the Irish EU-SLIC BALISE_CAPI questionnaires.

Field Staff return completed interviews to HQ on a weekly basis.

4. Processing of the data

When the data is received at HQ a number of preliminary checks are initiated to ensure that individual field staff members have achieved their quota of households and to give an indication of the completeness of the data. Missing households or large amounts of missing data will initiate a query to the relevant Field Staff member for an explanation.

Missing data and the need to impute for such missing values are features of all surveys and more so for complex household surveys. The EU-SILC is no different and missing data in this survey has been particularly concentrated in questions concerning Income, Tax & Social insurance payments, mortgage interest payments, the structural content of insurance premiums and farm income. In Irish EU-SILC programme each piece of missing data is individually estimated.

For estimating income values, information from the National Employment Authority is used. For estimating tax and social insurance contributions a ready-reckoner from the Irish Revenue Commissioners is used. For mortgage interest payments and insurance payments, ready-reckoners from the lenders or institutions are used. For estimating farm income, information is supplied by the National Farm Authority that is based on size, soil type and system used.

Annex II - Mechanisms for Non-response

Most surveys have some residual non-response even after careful design and follow-up of non-response. All methods for fixing up non-response are necessarily model-based. If we are to make any inferences about the non-respondents, we must assume that they are related to respondents in some way.

Dividing population members into two fixed strata of would–be respondents and would-be non-respondents is normally used when thinking about potential non-response bias. However, when adjusting for non-response that remains after all other measures have been taken, one needs a more elaborate set-up, letting the response or non-response unit *i* be a random variable. We define the random variable

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ responds} \\ 0 & \text{if unit } i \text{ does not respond} \end{cases}$$

After sampling, the realisations of the response indicator variable are known for the units selected in the sample. A value of y_i, (i.e. a characteristic associated with the *i* th unit in the population) is recorded if r_i, the realisation of R_i, is 1. The probability that a unit selected for the sample will respond,

$$\phi_i = P(R_i = 1)$$

is of course unknown but assumed positive.

Suppose that y_i is a response of interest and $\mathbf{x}_i$ is a vector of information known about unit *i* in the sample. Information used in the survey design is included in $\mathbf{x}_i$.

In survey sampling three types of missing are normally considered.

Missing Completely at Random

If ϕ_i does not depend on $\mathbf{x}_i$, y_i, or the survey design, the missing data are **missing completely at random** (MCAR). Such a situation occurs, if for example, someone at the laboratory drops a test tube containing the blood sample of one of the survey participants – there is no reason to think that the dropping of the rest tube has anything to do with white blood cell count. If data are MCA, the respondents are representative of the selected sample.

Missing data in the QNHS would be MCAR if the probability of non-response is completely unrelated to the region of Ireland, sex, age or any other variable measured for the sample and if the probability of non-response is unrelated to any variables about labour market status.

Missing at Random Given Covariates, or Ignorable Non-response

If ϕ_i depends on $\mathbf{x}_i$ but not on y_i, the data are **missing at random** (MAR); the non-response depends only on observed variables. We can successfully model the non-response, since we know the values of $\mathbf{x}_i$ for all sample units.

Persons in the QNHS would be missing at random if the probability of responding to the survey depends on, say sex and/or age – all known quantities – but does not vary with labour market status within each age/sex/region class.

This is sometimes termed **ignorable non-response**. Ignorable means that a model can explain the non-response mechanism and that the non-response can be ignored after the model accounts for it, not that the non-response can be completely ignored and complete-data methods used.

Non-ignorable Non-response

If the probability of non-response depends on the value of a response variable and cannot be completely explained by the $\mathbf{x}$'s, then the non-response is **non-ignorable**.

This is unlikely situation to be the situation for the QNHS, but could be for a Crime & Victimisation Survey where it is suspected that a person who has been victimised by crime is less likely to respond to the survey than a non-victim.

Weighting for Methods for non-response

In sample surveys the sampling weights (w_i) are used in calculating estimates of the population surveyed. The sampling weights are the reciprocals of the probabilities of selection, so an estimate of the population total (*Y*) is

$$\hat{Y} = \sum \frac{N}{n} y_i = \sum w_i y_i$$

Weights can also be used to adjust for non-response. Let Z_i be the indicator variable for the presence in the selected sample, with $P(Z_i = 1) = \pi_i$. If R_i is independent of Z_i, then the probability that unit *i* will be measured is

P(unit *i* selected in sample and responds) $= \pi_i \phi_i$.

The probability of responding, ϕ_i, is estimated for each unit in the sample, using auxiliary information that is know for all units in the selected sample. The final weight for a respondent is then $1/\pi_i \phi_i$. Weighting methods assume that the response probabilities can be estimated form variables known for all units; they assume MAR data.

Weighting- class adjustment

Sample weights w_i have been interpreted as the number of unit in the population represented by unit *i* in the sample. Weighting-class methods extend this approach to compensate for non-sampling errors. Variables known for all units in the selected sample are used to form weighting-adjustment classes, and it is hoped that respondents and non-respondents in the same weighting-adjustment class are similar. Weights of respondents in the weighting-adjustment class are increased so that the respondents represent the non-respondents' share of the population as well as their own.

For example, suppose the age is known for every member of the selected sample and that person *i* in the selected sample has the sampling weight $w_i = 1/\pi_i$. Then weighting classes can be formed by dividing the selected sample among different age classes as *Table 1* below shows.

We estimate the response probability for each class by

$$\phi_c = \frac{\text{sum of weights for respondents in class c}}{\text{sum of weights for selected sample in class c}}$$

Then the sampling eight for each respondent class *c* is multiplied by $1/\hat{\phi}_c$, the weight factor in *Table 1*. The weight of each respondent with aged between 15 and 24, for example, is multiplied by 1.622. Since there was no non-response in the over 65group, their weights are unchanged.

Table A: Illustration of Weighting-Class Adjustment Factors

	Age					
	15-24	25-34	35-44	45-64	65+	Total
Sample size	202	220	180	195	203	1,000
Respondents	124	187	162	187	203	863
Sum of weights for sample	30,322	33,013	27,046	29,272	30,451	150,104
Sum of weights for respondents	18,693	28,143	24,371	28,138	30,451	
$\hat{\phi}_c$,	0.6165	0.8525	0.9011	0.9613	1.0000	
Weight factor	1.622	1.173	1.110	1.040	1.000	

The probability of response is assumed to be the same within each weighting class, with the implication that within a weighting class, the probability of response does not depend on *y*. As mentioned earlier, weighting-class methods assume MAR data. The weight for a respondent in the weighting class *c* is $1/\pi_i\phi_c$.

To estimate the population total using weighting-class adjustments, let $x_{ci} = 1$ if unit *i* is in class *c*, and 0 otherwise. Then let the new weight for respondent *i* be

$$\tilde{w}_i = \sum_c \frac{w_i x_{ci}}{\hat{\phi}_c},$$

where w_i is the sampling weight for unit *i*; $\tilde{w}_i = w_i / \phi_c$ if unit *i* is in class *c*. Assign $\tilde{w}_i = 0$ if unit *i* is a non-respondent. Then,

$$\hat{t}_{wc} = \sum_{i \in S} \tilde{W}_i y_i$$

and

$$\hat{\bar{y}}_{wc} = \frac{\hat{t}_{wc}}{\sum_{i \in S} \tilde{w}_i}.$$

Post-stratification

Post-stratification is similar to weighting-class adjustment, except that population counts are used to adjust the weights. Suppose a Simple Random Sample (SRS) is taken. After the sample is collected, units are grouped in *H* different post-strata, usually based on demographic variables such as age and sex. The population has N_h units in post-stratum *h*; of these n_h were selected for the sample and n_{hR} responded. The post-stratified estimator for $\bar{y}_U$ is

$$\bar{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{hR};$$

the weighting-class estimator for $\bar{y}_U$, if the weighting classes are post-strata, is

$$\bar{y}_{wc} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_{hR}.$$

The two estimators are similar in form; the only difference is that in post stratification the N_h are known, whereas in weighting-class adjustments the N_h are unknown and estimated by Nn_h/n.

Post stratification using weights

In general survey design, the sum of the weights in subgroup *h* is supposed to estimate the population count N_h for that subgroup. Post stratification uses the ratio estimator within each subgroup to adjust by the true population count.

Let $x_{hi} = 1$ if unit *i* is a respondent in post-stratum *h*, and *0* otherwise. Then let

$$\overset{*}{W}_i = \sum_{h=1}^{H} w_i x_{hi} \frac{N_h}{\sum_{j \in S} w_j x_{hj}} .$$

Using modified weights,

$$\sum_{i \in S} \overset{*}{w} x_{hi} = N_h$$

and the post-stratified estimator of the population total is

$$\hat{t}_{post} = \sum_{i \notin S} \overset{*}{w}_i y_i .$$

Post-stratification can adjust for under coverage as well as non-response if the population count N_h includes individuals not in the sample frame for the survey.

Raking adjustments

Raking is a post-stratification method that can be used when post-strata are formed using more that one variable, but only the marginal population totals are known.

Consider the following table of sums of weights from a sample; each entry in the table is the sum of the sampling weights for persons in the sample falling in that classification

	NACE					
	A-B	*C-E*	*F*	*G*	*H*	*Sum of weights*
Male	300	1,200	60	30	30	1,620
Female	150	1,080	90	30	30	1,380
Sum of weights	450	2,280	150	60	60	3,000

Now suppose we know the true population counts for the marginal totals. We know that the population has 1,510 women and 1,490 men, 600 A-Bs, 2,120 C-Es, 150 Fs, 100 Gs and 30 persons in the H category. The population counts for each cell in the table, however, are unknown; we do not know the number of A-B females in the population and cannot assume independence. Raking allows us to adjust the weights so that the sums of weights in the margins equal the population counts.

First adjust the rows. Multiply each entry by (true row population)/(estimated row population). Multiplying the cells in the female row by 1,510/1,620 and the cells in the male row by 1,490/1,380 results in the following table.

	NACE					
	A-B	*C-E*	*F*	*G*	*H*	*Sum of weights*
Male	279.63	1,118.52	55.93	27.96	27.96	1,510
Female	161.96	1,166.09	97.17	32.39	32.39	1,490

Sum of weights	441.59	2,284.61	153.10	60.35	60.35	3,000

The row totals are fine now, but the column totals do not yet equal the population totals. Repeat the same procedure with the columns in the new table. The entries in the first column are each multiplied by 600/441.59. The following table results:

	NACE					
	A-B	*C-E*	*F*	*G*	*H*	*Sum of weights*
Male	379.94	1,037.93	54.79	46.33	13.90	1,532.90
Female	220.06	1,082.07	95.21	53.67	16.10	1,467.10
Sum of weights	600.00	2,120.00	150.00	100.00	30.00	3,000.00

But this has thrown the row totals off again. Repeat the procedure until both the row and column totals equal the population counts. The procedure converges as long as all cell counts are positive. IN this example, the final table of adjusted counts is

	NACE					
	A-B	*C-E*	*F*	*G*	*H*	*Sum of weights*
Male	375.59	1,021.47	53.72	45.56	13.67	1,510
Female	224.41	1,098.53	96.28	54.44	16.33	1,490
Sum of weights	600.00	2,120.00	150.00	100.00	30.00	3,000

The entries in the last table may be better estimate of the cell populations (that is, with smaller variances) than the original weighted estimates, simply because they use more information about the population. The weighting –adjustment factor for each C-E male in the sample is 1098.53/1080; the weight of each white male is increase a little to adjust for non-response and under-coverage. Likewise, the weights of C-E females are decreased because they are over-represented in the sample.

The assumptions for *raking* are the same for post-stratification with the additional assumption that the response probabilities depend only on the row and column and not on the particular cell. If the sample sizes in each cell are large enough, the raking estimator is approximately unbiased.

Annex III - Contingency tables

In statistics, **contingency tables** are used to record and analyse the relationship between two or more variables, most usually categorical variables

Suppose that we have two variables, sex (male or female) and handedness (right-handed or left-handed). We observe the values of both variables in a random sample of 100 people. Then a contingency table can be used to express the relationship between these two variables, as follows:

	right-handed	left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

The figures in the right-hand column and the bottom row are called **marginal totals** and the figure in the bottom right-hand corner is the **grand total**.

The table allows us to see at a glance that the proportion of men who are right-handed is about the same as the proportion of women who are. However the two proportions are not identical, and the statistical significance of the difference between them can be tested with a Pearson's chi-square test, a G-test or Fishers's exact test, provided the entries in the table represent a random sample from the population contemplated in the null hypothesis. If the proportions of individuals in the different columns varies between rows (and, therefore, vice versa) we say that the table shows *contingency* between the two variables. If there is no contingency, we say that the two variables are *independent*.

The example above is for the simplest kind of contingency table, in which each variable has only two levels; this is called a 2 x 2 contingency table. In principle, any number of rows and columns may be used. There may also be more than two variables, but higher order contingency tables are hard to represent on paper. The relationship between ordinal variables, or between ordinal and categorical variables, may also be represented in contingency tables, though this is less often done since the distributions of ordinal variables can be summarised efficiently by the median.

The degree of association between the two variables can be assessed by a number of coefficients: the simplest is the *phi* coefficient defined by

$$\varphi = \sqrt{\frac{\chi^2}{N}}$$

where χ^2 is derived from the Pearson test, and *N* is the grand total number of observations. φ varies from 0 (corresponding to no association between the variables) to 1 (complete association). This coefficient can only be used for 2 x 2 tables. Alternatives include the **tetrachoric correlation coefficient** (also only useful

for 2 x 2 tables), the **contingency coefficient** *C* and Cramer's ***V***. *C* suffers from the disadvantage that it does not reach a maximum of 1 with complete association in asymmetrical tables (those where the numbers of row and columns are not equal). The tetrachoric correlation coefficient is essentially the Pearson product-moment correlation coefficient between the row and column variables, their values for each observation being taken as 0 or 1 depending on the category it falls into. The formulae for the other coefficients are:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

k being the number of rows or the number of columns, whichever is less. *C* can be adjusted so it reaches a maximum of 1 when there is complete association in a table of any number of rows and columns by dividing it by $\sqrt{\frac{(k-1)}{k}}$.

Pearson's chi-square test

Pearson's chi-square test χ^2 is one of a variety of chi-square tests – statistical procedures whose results are evaluated by reference to the chi-square distribution. It tests a null hypothesis that the relative frequencies of occurrence of observed events follow a specified frequency distribution. The events must be mutually exclusive. One of the simplest examples is the hypothesis that an ordinary six-sided die is "fair", i.e., each of the six possible outcomes occurs equally often. Chi-square is calculated by finding the difference between the observed and theoretical (expected) frequencies for each event, squaring these differences, dividing each by the theoretical frequency, and taking the sum of the results.

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

where:

O_i = an observed frequency

E_i = an expected (theoretical) frequency, asserted by the null hypothesis

For example, to test the hypothesis that a random sample of 100 people has been drawn from a population in which men and women are equal in frequency, the observed number of men and women would be compared to the theoretical frequencies of 50 men and 50 women. If there were 45 men in the sample and 55 women, then

$$\chi^2 = \frac{(45-50)^2}{50} + \frac{(55-50)^2}{50} = 1$$

If the null hypothesis is true (ie men and women are chosen with equal probability in the sample), the test statistic will be drawn from a chi-square distribution with one degree of freedom. One might expect there to be two degrees of freedom, one for the male count and one for the female. However, in this case there is only one degree of freedom because the male and female count are constrained to have a sum of 100 (the sample size), and this constraint reduces the number of degrees of freedom by one. Alternatively, if the male count is known the female count is determined, and vice-versa.

Consultation of the chi-square distribution for 1 degree of freedom shows that the probability of observing this difference (or a more extreme difference than this) if men and women are equally numerous in the population is approximately 0.3. This probability is higher than conventional criteria for statistical significance, so normally we would not reject the null hypothesis that the number of men in the population is the same as the number of women.

The approximation to the chi-square distribution breaks down if expected frequencies are too low. It will normally acceptable so long as no more than 10% of

the events have expected frequencies below 5. Where there is only 1 degree of freedom, the approximation is not reliable if expected frequencies are below 10. In this case, a better approximation can be had by reducing the absolute value of each difference between observed and expected frequencies by 0.5 before squaring; this is called Yates' correction.

Pearson's chi-square is used to assess two types of comparison: tests of goodness of fit and tests of independence. A test of goodness of fit establishes whether or not an observed frequency distribution differs from a theoretical distribution. A test of independence assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other – for example, whether people from different regions differ in the frequency with which they report that they support a political candidate.

Pearson's chi-square is the original and most widely-used chi-square test.

The null distribution of the Pearson statistic is only approximated as a chi-square distribution. This approximation arises as the true distribution, under the null hypothesis, if the expected value is given by a multinomial distribution. For large sample sizes, the central limit theorem says this distribution tends toward a certain multivariate normal distribution. In the special case where there are only two cells in the table, the expected values follow a binomial distribution,

$$E \quad = \quad {}^{d}B(n, p),$$

where

p = probability, under the null hypothesis,

n = number of observations in the sample.

In the above example the hypothesised probability of a male observation is .5, with 100 samples. Thus we expect to observe 50 males.

When comparing the Pearson test statistic against a chi-squared distribution, the above binomial distribution is approximated as a Gaussian (normal) distribution:

$$Bin(n, p) \approx \quad {}^{d}N(np, np(1-p)).$$

Let O_1 be the number of observations from the sample that are in the first cell. The Pearson test statistic can be expressed as

$$\frac{(O_1 - np)^2}{np} + \frac{((n - O_1) - n(1-p))^2}{n(1-p)}$$

which can in turn be expressed as

$$\left(\frac{O_1 - np}{\sqrt{np(1-p)}}\right)^2$$

By the normal approximation to a binomial this is the square of one standard normal variate, and hence is distributed as chi-square with *1* degree of freedom.

In the general case where there are *k* cells in the contingency table, the Normal approximation results in a sum of $k - 1$ standard normal variates, and is thus distributed as chi-square with $k - 1$ degrees of freedom (ignoring the issue of dependence between the variates):

In cases whereby the expected value, *E*, is found to be small (indicating either a small underlying population probability, or a small number of observations), the normal approximation of the multinomial distribution can fail, and in such cases it is found to be more appropriate to use the G-test, a Likelihood ratio-based test statistic. Where the total sample size is small, it is necessary to use an appropriate exact test, typically either the binomial test or (for contingency tables) Fishers exact test..

A more complicated, but more widely used form of Pearson's chi-square test arises in the case where the null hypothesis of interest includes unknown parameters . For instance we may wish to test whether some data follows a normal distribution but without specifying a mean or variance. In this situation the unknown parameters need to be estimated by the data, typically by maximum likelihood estimation, and these estimates are then used to calculate the expected values in the Pearson statistic. It is commonly stated that the degrees of freedom for the chi-square distribution of the statistic are then $k - 1 - r$, where *r* is the number of unknown parameters. This result is valid when the original data was Multinomial and hence the estimated parameters are efficient for minimising the chi-square statistic. More generally however, when maximum likelihood estimation does not coincide with minimum chi-square estimation, the distribution will lie somewhere between a chi-square distribution with $k - r - 1$ and $k - 1$ degrees of freedom (See for instance Chernoff and Lehmann 1954).

G-tests

In statistics, **G-tests** are likelihood-ratio or maximum likelihood statistical significance tests that are increasingly being used in situations where chi-square tests were previously recommended.

The commonly used chi-squared tests for goodness of fit to a distribution and for independence in contingency tables are in fact approximations of the log-likelihood ratio on which the G-tests are based. This approximation was developed by Karl Pearson because at the time it was unduly laborious to calculate log-likelihood ratios. With the advent of electronic calculators and personal computers, this is no longer a problem. G-tests are coming into increasing use

The general formula for Pearson's chi-squared test statistic is

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the frequency observed in a cell, E_i is the frequency expected on the null hypothesis, and the sum is taken across all cells. The corresponding general formula for *G* is

$$G = 2\sum_i O_i . \ln\left(\frac{O_i}{E_i}\right)$$

where *ln* denotes the natural logarithm (log to the base *e*) and the sum is again taken over all cells. The value of *G* can also be expressed in terms of mutual information of the contingency table or as the difference of the entropy of the contingency table and the entropy of the row and column sums

$$G = 2\sum_{ij} k_{ij} \left[H(k_{ij}) - H(k_{i.}) - H(k_{.j}) \right]$$

where the entropy of some observed distribution *q* is defined as

$$H(q_i) = -\sum_i q_i \log q_i$$

and

$$\pi_{ij} = \frac{k_{ij}}{\sum_i jk_{ij}},\ \pi_{i.} = \frac{\sum_j k_{ij}}{\sum_{ij} jk_{ij}},\ \text{and}\ \pi_{.j} = \frac{\sum_i k_{ij}}{\sum_{ij} jk_{ij}}.$$

.
It can also be shown that the inverse document frequency weighting commonly used for text retrieval is an approximation of *G* where the row sum for the query is much

smaller than the row sum for the remainder of the corpus. Similarly, the result of Bayesian inference applied to a choice of single multinomial distribution for all rows of the contingency table taken together versus the more general alternative of a separate multinomial per row produces results very similar to the *G* statistic.

Given the null hypothesis that the observed frequencies result from random sampling from a distribution with the given expected frequencies, the distribution of *G* is approximately that of chi-squared, with the same number of degrees of freedom as in the corresponding chi-squared test.

For samples of a reasonable size, the *G*-test and the chi-squared test will lead to the same conclusions. However, the approximation to the theoretical chi-square distribution for the G-test is better than for the Pearson chi-squared tests in cases where for any cell $|O_i - E_i| > E_i$, and in any such case the *G*-test should always be used.

For very small samples the multinomial test for goodness of fit, and Fisher's exact test for contingency tables, or even Bayesian hypothesis selection are preferable to either the chi-squared test or the *G*-test.

Fisher's exact test

Fisher's exact test is a statistical significance test used in the analysis of categorical data where sample sizes are small. It is named after its inventor, R. A. Fisher, and is one of a class of exact tests. Fisher devised the test following a comment from Muriel Bristol, who claimed to be able to detect whether the tea or the milk was added first to her cup.

The test is used to examine the significance of the association between two variables in a 2 x 2 contingency table. With large samples, a chi-squared test can be used in this situation. However, this test is not suitable when the "expected values" in any of the cells of the table is below 10 and there is only one degree of freedom: the sampling distribution of the test statistic that is calculated is only approximately equal to the theoretical chi-squared distribution, and the approximation is inadequate in these conditions (which
arise when sample sizes are small, or the data are very unequally distributed among the cells of the table). The Fisher test is, as its name states, exact, and it can therefore be used regardless of the sample characteristics. It becomes difficult to calculate with large samples or well-balanced tables, but fortunately these are exactly the conditions where the chi-square test is available.

The need for the Fisher test arises when we have data that are divided into two categories in two separate ways. For example, a sample of teenagers might be divided into male and female on the one hand, and those that are and are not currently dieting on the other. We hypothesise, perhaps, that the proportion of dieting individuals is higher among the women than among the men, and we want to test whether any difference of proportions that we observe is significant. The data might look like this:

	Men	*Women*	*Total*
Dieting	1	9	10
Not dieting	11	3	14
Totals	12	12	24

These data would not be suitable for analysis by a chi-squared test, because the expected values in the table are all below 10, and in a 2 × 2 contingency table, the number of degrees of freedom is always 1. The question we ask about these data is: knowing that 10 of these 24 teenagers are dieters, what is the probability that these 10 dieters would be so unevenly distributed between the girls and the boys? If we were to choose 10 of the teenagers at random, what is the probability that 9 of them would be among the 12 girls, and only 1 from among the 12 boys? Before we proceed with the Fisher test, we first introduce some notation. We represent the cells by the letters *a, b, c* and *d*, call the totals across rows and columns *marginal totals*, and represent the grand total by *n*. So the table now looks like this:

	Men	*Women*	*Total*

Dieting	a	b	a+b
Not dieting	c	d	c+d
Totals	a+c	b+d	n

Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

$$= \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

where the symbol ! indicates the factorial, i.e. 1 multiplied by 2 multiplied by 3 etc, up to the number whose factorial is required. This formula gives the exact probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that the odds ratio between dieter and non-dieter among men and women equals to 1 in the population from which our sample was drawn. Fisher showed that we could only deal with cases where the marginal totals are the same as in the observed table. In the example, there are 11 such cases. Of these only one is more extreme in the same direction as our data; it looks like this:

	Men	*Women*	*Total*
Dieting	0	10	10
Not dieting	12	2	14
Totals	12	12	24

In order to calculate the significance of the observed data, i.e. the total probability of observing data as extreme or more extreme if the null hypothesis is true, we have to calculate the p values for both these tables, and add them together. This gives a one-tailed test; for a two-tailed test we must also consider tables that are equally extreme but in the opposite direction. Unlike most statistical tests, it is not always the case that the two-tailed significance level is exactly twice the one-tailed significance level. In the
example above, the one-tailed significance level is 0.0014 and the two-tailed significance level is twice this, since this problem is symmetrical (same number of boys as girls). Because the calculation of Fisher's exact test involves permuting the observed cell frequencies it is referred to as a permutation test, one of a broad class of such tests. Calculating significance values for the Fisher exact test with a calculator is slow and requires care because the factorial terms quickly become very large, and with larger samples, the number of possible tables more extreme than that observed quickly becomes substantial. Even for small samples (which fortunately is where the test is usually needed), the calculations are tedious, but published tables

are available; they are bulky, because the grand total and two of the four cell sizes have to be specified. Given these data, the table then gives the criterial value of the third cell size for specified significance levels. The observed table may have to be re-arranged (for example by rearranging the rows or the columns) to make it compatible with the way the significance levels are tabulated. Most modern statistical packages will calculate the significance of Fisher tests, in some cases even where the chi-squared approximation would also be acceptable.

Pearson product-moment correlation coefficient

In statistics, the *Pearson product-moment correlation coefficient* (sometimes known as the PMCC) (*r*) is a measure of how well a linear equation describes the relation between two variables *X* and *Y* measured on the same object or organism. It is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom:

$$r = \frac{\sum z_x z_y}{N-1}$$

Note that this formula assumes that the standard deviations on which the *Z* scores are based are calculated using *N-1* in the denominator.

The result obtained is equivalent to dividing the covariance between the two variables by the product of their standard deviations. In general the correlation coefficient is one of the two square roots (either positive or negative) of the coefficient of determination (r^2), which is the ratio of explained variation to total variation:

$$r^2 = \frac{\sum (Y' - \bar{Y})^2}{\sum (Y - \bar{Y})^2}$$

where:

Y = a score on a random variable *Y*

Y' = corresponding predicted value of *Y*, given the correlation of *X* and *Y* and the value of *X*

$\bar{Y}$ = sample mean of *Y* (i.e., the mean of a finite number of independent observed realisations of *Y*, not to be confused with the expected value of *Y*)

The correlation coefficient adds a sign to show the direction of the relationship. The formula for the Pearson coefficient conforms to this definition, and applies when the relationship is linear.

The coefficient ranges from *−1* to *1*. A value of *1* shows that a linear equation describes the relationship perfectly and positively, with all data points lying on the same line and with *Y* increasing with *X*. A score of *−1* shows that all data points lie on a single line but that *Y* increases as *X* decreases. A value of *0* shows that a linear model is inappropriate – that there is no linear relationship between the variables.

The Pearson coefficient is a statistic which estimates the correlation of the two given random variables.

The linear equation that best describes the relationship between *X* and *Y* can be found by linear regression. This equation can be used to "predict" the value of one measurement from knowledge of the other. That is, for each value of *X* the equation

calculates a value which is the best estimate of the values of *Y* corresponding the specific value of *X*. We denote this predicted variable by Y.

Any value of *Y* can therefore be defined as the sum of *Y′* and the difference between *Y* and *Y′*:

$$Y = Y + (Y\text{-}Y')$$

The variance of *Y* is equal to the sum of the variance of the two components of Y:

$$s_y^2 = s_{y'}^2 + s_{y.x}^2$$

Since the coefficient of determination implies that $s_{y.x}^2 = s_y^2 (1 - r^2)$ we can derive the identity

$$r^2 = \frac{s_{y'}^2}{s_y^2}$$

The square of *r* is conventionally used as a measure of the association between *X* and *Y*. For example, if the coefficient is 0.90, then 81% of the variance of *Y* can be "accounted for" by changes in *X* and the linear relationship between *X* and *Y*.

References

Cassel, C.M., Granström, F., Lundquist, P and J. Selén (1997). Cumulating Data from Household Budget Survey. Some Results for Model Based Calibration Techniques Applied to Swedish Data. *Report financed the by European Communities, LOT 23.*

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63,615-620.

Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics*, 30, 109-126.

Deaton, A. and J. Muellbauer (1980). An almost ideal demand system. *American Economic Review*, 70, 312-126.

Deming, W.E. and F. F. Stephan (1940). On a least squares adjustment of a sampled frequency when the expected marginals are known. *The Annals of Mathematical Statistics, 11, 427- 444*

Deville, J.-C. (2000) Generalized Calibration and Application to Weighting for Non-response. Proceedings of the 14th Compstat Symposium in Computational Statistics, Utrecht, Physica-Verlag.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Harms, T. and Dushesne, P. (2006).On Calibration Estimates for Quantiles. *Survey Methodology,* June 2006.

Cryer, J.D., Miller, R.B. (1991) Statistics for Business: Data Analysis and Modelling. *PWS-KENT.*

Särndal, C.-E, Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer.