Bucharest, 13 December 2023

The Use of R in Official Statistics - uRos2023

Assessing coherence between estimated distributions in R

Marcello D'Orazio
Istat | Directorate for Methodology and Statistical Process Design

Outline

- Coherence of statistics
- Assessing coherence between estimated distributions: **categorical** variables
- Assessing coherence between estimated distributions: **continuous** variables

Coherence of statistics

- **Coherence**, jointly with **comparability**, is part of the ESS definition of **quality of statistics**.

- **Coherence**:
  "assessing the extent to which the outputs from different statistical processes have the potential to be reliably used in combination"

  Incoherence and **non-comparability** can affect statistics originating from different sources.
  Causes may be:
  - Differences in concepts (a household could be defined in a number of ways…)
  - Differences in methods (e.g. employment estimated from a household survey Vs. employment estimated from administrative data)

  ESS, *Handbook for Quality and Metadata Report, 2021 re-edition*

- Assessing coherence becomes crucial in modern statistical production processes involving integration of data from different sources (exploitation of variables shared by the sources)

Coherence: ESS SIMS

SIMS	Concept Name	Definition	Summary Guidelines
S.15.3	Coherence- cross domain	The extent to which statistics are reconcilable with those obtained through other data sources or statistical domains.	An analysis of incoherence should be provided, where this is an issue of importance. Reporting under 15.3 is for coherence problems that are not reported under 15.3.1, 15.3.2 or 15.4
S.15.3.1	Coherence - subannual and annual statistics (P)	The extent to which statistics of different frequencies are reconcilable.	*For producer reports only.* Coherence between subannual and annual statistical outputs is a natural expectation but the statistical processes producing them are often quite different. Compare subannual and annual estimates and, eventually, describe reasons for lack of coherence between subannual and annual statistical outputs.
S.15.3.2	Coherence- National Accounts (P)	The extent to which statistics are reconcilable with National Accounts.	*For producer reports only.* Where relevant, the results of comparisons with the National Account framework and

“Where possible, a quantitative analysis of any lack of coherence should be presented”

Coherence Assessment

Currently assessment is based on comparison of estimates:

- Occurrence of given categories of a categorical variable
- Average, totals, percentiles for continuous variables

It is preferable to assess coherence between estimated marginal distributions

Different scenarios depending on the type of data source:

- Estimates from two independent random samples (complex sampling design)
- Estimate from a sample survey and an estimate from a nonprobabilistic data source (non-prob. sample, admin. data, big data, etc.)

Is it available a "reference" estimate? I.e. an estimate considered reliable and therefore the reference one

Coherence Between distributions: categorical variables (1/3)

Category	Source_1	Source_2
1	$\hat{p}_{11}$	$\hat{p}_{12}$
2	$\hat{p}_{21}$	$\hat{p}_{22}$
…	…	…
j	$\hat{p}_{j1}$	$\hat{p}_{j2}$
…	…	…
J	$\hat{p}_{J1}$	$\hat{p}_{J2}$
Total	1.00	1.00

$$\hat{p}_{ji} = \hat{N}_{ji}/\hat{N}_i\,, \qquad i = 1,2$$

In probabilistic sample surveys:

$$\hat{p}_{ji} = \sum_{k=1}^{n_i} w_{ki} I(y_{ki} = j)$$

Total Variation Distance (TVD) $\Delta_{12} = \frac{1}{2}\sum_{j=1}^{J} \left|\hat{p}_{j1} - \hat{p}_{j2}\right|$ $\quad 0 \leq \Delta_{12} \leq 1$

Overlapping coefficient $O_{12} = 1 - \Delta_{12}$ $\quad 0 \leq O_{12} \leq 1$

Bhattacharyya coefficient $B_{12} = \sum_{j=1}^{J} \sqrt{\hat{p}_{j1} \times \hat{p}_{j2}}$ $\quad 0 \leq B_{12} \leq 1$

Hellinger distance $d_{H,12} = \sqrt{1 - B_{12}}$ $\quad 0 \leq d_{H,12} \leq 1$

$$d_{H,AB}^2 \leq \Delta_{AB} \leq d_{H,AB}\left(\sqrt{2}\right)$$

Rule of thumbs: if $\hat{p}_{j2}$ is the reference:

$\hat{p}_{j1}$ is «close» to $\hat{p}_{j2}$ when $\Delta_{12} \leq 0.03$ (Agresti, 2002)

$\hat{p}_{j1}$ is «close» to $\hat{p}_{j2}$ when $d_{H,12} \leq 0.05$ (??) $\quad d_{H,12} \leq 0.0212$

Coherence Between distributions: categorical variables (2/3)

New R function `comp.tables()`, derived from `comp.prop()` in **StatMatch** (D'Orazio, 2022)

```
> data(samp.A, package = "StatMatch")
> data(samp.B, package = "StatMatch")

> t.edu.A <- xtabs(ww~edu7, data=samp.A)
> t.edu.B <- xtabs(ww~edu7, data=samp.B)
> t.edu.B
edu7
         0          1          2          3          4          5          6
 149580.43  997271.57 1604170.80 1687398.23  141106.95  564485.98   13568.23

> comp.tables(p1 = t.edu.A, p2 = t.edu.B,
+             ref = TRUE) # t.edu.B is the reference one
       tvd    overlap      Bhatt       Hell
0.01048456 0.98951544 0.99986854 0.01146559
```

Coherence Between distributions: categorical variables (3/3)

Estimates from two independent sample surveys referred to the same target population and **no reference**

- Reference estimate obtained by «pooling» (Sarndal et al 1992; Korn & Graubard, 1999):

$$\hat{p}_{j,r} = \lambda_1 \hat{p}_{j1} + (1 - \lambda_1)\hat{p}_{j2} \qquad \lambda_1 = \frac{n_1}{n_1 + n_2}$$

- Alternative ways for estimating λ_1 (O'Muirchertaigh & Pedlow, 2002)

$$\lambda_1 = \frac{n_1/d_{w1}}{n_1/d_{w1}+n_2/d_{w2}}, \qquad d_{wi} = 1 + CV_{w_i}^2$$

- These and other options implemented in the a new R function `opt.lambda()`

```
> data(samp.A, package = "StatMatch")
> data(samp.B, package = "StatMatch")
> opt.lambda(w1 = samp.A$ww, w2 = samp.B$ww)
$summaries.w
                 s1           s2
n      3.009000e+03 6.686000e+03
N      5.094952e+06 5.157582e+06
Nc     1.006146e+00 9.939283e-01
mean.w 1.693238e+03 7.714003e+02
sd.w   1.203468e+03 5.339756e+02
CV.w   7.107498e-01 6.922160e-01
deff.w 1.505165e+00 1.479163e+00

$lambdas
            s1        s2       tot
kg1  0.3085334 0.6891416 0.9976750
kg2a 0.3122738 0.6854466 0.9977204
kg2b 0.3066486 0.6933514 1.0000000
kg3  0.3103662 0.6896338 1.0000000
omp  0.3066486 0.6933514 1.0000000
```


Istat

Coherence Between distributions: continuous variables

Two approximate approaches:

- **Comparison of percentiles (Q-Q)**
- **Categorization and estimation of indicators for categorical variables** (TVD, Hellinger’s distance, etc.)

Istat

Coherence Between distributions: percentiles of continuous variables (1/2)

An interesting expression (typically $p = 0.10$):

$$Q_p = Q_{0.5} + \frac{1}{2} IQR \times \frac{Q_{1-p} - Q_p}{IQR} \times \left(\frac{Q_p + Q_{1-p} - 2Q_{0.5}}{Q_{1-p} - Q_p} - 1 \right) \qquad IQR = Q_{0.75} - Q_{0.25}$$

Median (location)

IQR (scale)

Shape index

Skewness (shape)
Bowley's index with *p*=0.25

Q_p should estimated using survey weights, when available (see e.g. Korn & Graubard, 1999) -> `wtd.qs()`

In alternative compare percentiles (quartiles; quintiles, deciles,…)

$$\hat{Q}_{pi} - \hat{Q}_{pr} \qquad \frac{\left(\hat{Q}_{pi} - \hat{Q}_{pr}\right)}{\hat{Q}_{pr}} \qquad i = 1,2; \qquad p = 0.25, 0.50, 0.75$$ in the case of quartiles, and so on…

If there are no reference $\hat{Q}_{pr}$ and the data come from two independent sample surveys referred to the same target population, then $\hat{Q}_{pr}$ should be estimated on the concatenated sample with weights

$$\tilde{w}_{ki} = \lambda_i w_{ki}, \qquad k = 1,2,\dots,n_i, \qquad i = 1,2$$

Coherence Between distributions: percentiles of continuous variables (2/2)

The Median, IQR, shape and skewness based on Quantiles are returned by the R function **smrs()**

```
> smrs(x=samp.A$n.income, weights = samp.A$ww, p = 0.10)
$summary
       Min        P10         Q1     Median       Mean         Q3        P90        Max
-15000.000      0.000   3977.326  12497.762  13978.449  19825.173  28185.414 276750.000

$qq.based
           p          IQR         shape     skewness
1.000000e-01 1.584785e+04 1.778501e+00 1.131752e-01
```

While comparison of quantiles is performed by the R function **comp.quantiles()**

```
> comp.quantiles(x1 = samp.A$age, x2 = samp.B$age, w1 = samp.A$ww, w2 = samp.B$ww,
+                pctp = seq(0.1,0.9,0.1), ref = TRUE)

  Pct qqs.1 qqs.2 qqs.ref diff    rel.diff
1 P10    24    25      25   -1 -0.04000000
2 P20    32    33      33   -1 -0.03030303
...
8 P80    68    68      68    0  0.00000000
9 P90    77    77      77    0  0.00000000
```

Coherence Between distributions: categorize continuous variables (1/4)

Discretization

Freedman & Diaconis (1981) rule for histogram bin width:

$$b = 2 \times \frac{IQR}{\sqrt[3]{n_0}}$$

No. of bins:

$$m = \left[\frac{x_u - x_l}{b}\right] + 1 \qquad x_l \leq x_{min} \quad x_u \geq x_{max}$$

Instead of min and max it is possible to consider bounds for detection of outliers (see functions `boxB()` or `LocScaleB()` in **univOutl**)

$$n_0 = \min(n_1, n_2)$$

In case of sample surveys, replace n_i with n_i/d_{wi}

IQR should be estimated on the reference data source (using weights if data come from a prob. sample survey)

When data are from two independent sample surveys and there's NOT a reference then concatenate the samples and use **new weights**:

$$\widetilde{w}_{ki} = \lambda_i w_{ki}, \qquad k = 1,2,\ldots,n_i, \qquad i = 1,2$$

to estimate IQR

Coherence Between distributions: categorize continuous variables (2/4)

In R two new functions:

```
wtd.qs (x, w, prb, ties=FALSE)
```

to estimate quantiles using survey weights (considers possibility of tied values)

(many alternative functions exist in R packages with different estimation methods)

```
hist.bks(x, w = NULL, neff = NULL,
         robust=0,...)
```

to get the breaks to categorize **x**

In case of sample surveys replace n_i with n_i/d_{wi}

IQR should be estimated on the reference data source (using weights if data come from a prob. sample survey)

When data are from two independent sample surveys and there's NOT a reference then concatenate the samples and use new weights:

$$\tilde{w}_{ki} = \lambda_i w_{ki}, \qquad k = 1,2,\ldots,n_i, \qquad i = 1,2$$

to estimate IQR

Coherence Between distributions: categorize continuous variables (3/4)

```
> source("wtd.qs.R")
> source("hist.bks.R")

> bk.0 <- hist.bks(x = samp.A$n.income, w = samp.A$ww, neff = NULL, robust = 0)
n and eff_n:  3009 1999.339
width:  2515.966
min & max:  -15000 276750
mod low & up bounds:  -15051.04 276801
bins:  116
```

Coherence Between distributions: categorize continuous variables (4/4)

Categorization based on histograms permits estimating the **density** (Bellhouse & Stafford, 1999):

$$\hat{f}_B(x) = \frac{1}{h_B}\sum_{l=1}^{m} \hat{p}_l K_B\left(\frac{x-\tilde{x}_l}{h_B}\right)$$

h_B: bandwidth (rule of thumb $h_B = b/1.25$)
$\hat{p}_l$: estimated prop. of obs. (weighted) in the bin l
$K_B(\cdot)$: kernel function
$\tilde{x}_l$: midpoint of the bin l

```
> bk.0 <- hist.bks(x = samp.A$n.income, w = samp.A$ww, neff = NULL, robust = 1)
> oo <- discr.sum(x=samp.A$n.income, w=samp.A$ww, breaks = bk.0$breaks, density = TRUE)
> head(oo$binned.sum, 4)
                        cxx      Freq      relFreq      low.b   midpoint       up.b
1 [-1.51e+04,-1.26e+04] 2002.5312 3.930422e-04 -15147.506 -13889.523 -12631.5395
2 (-1.26e+04,-1.01e+04]    0.0000 0.000000e+00 -12631.539 -11373.556 -10115.5733
3  (-1.01e+04,-7.6e+03]  401.9409 7.889002e-05 -10115.573  -8857.590  -7599.6072
4  (-7.6e+03,-5.08e+03]  649.5610 1.274911e-04  -7599.607  -6341.624  -5083.6411

> head(oo$est.dens, 4)
       x         dens
1 -15000 6.705574e-08
2  -9000 3.036892e-08
3  -7000 4.574928e-08
4  -1672 1.615645e-05
```

Coherence Between distributions: Future developments

Future:

- Introduce comparison of estimated empirical cumulative distribution function (P-P) for continuous variables
- evaluate whether to create a new R package

Repository with R code and supporting material

https://github.com/marcellodo/coherenceD

Thank You

Marcello D’Orazio | marcello.dorazio@istat.it

Istat | Istituto Nazionale di Statistica

Main References

Agresti A. (2002) *Categorical Data Analysis. Second Edition*. Wiley, New York.

Bellhouse D.R., Stafford J. E. (1999) "Density Estimation From Complex Surveys". *Statistica Sinica*, 9, pp. 407-424

D'Orazio M (2022). *StatMatch: Statistical Matching or Data Fusion.* R package version 1.4.1, https://CRAN.R-project.org/package=StatMatch

D'Orazio M (2022). *univOutl: Detection of Univariate Outliers*. R package version 0.4, https://CRAN.R-project.org/package=univOutl

European Statistical System (ESS) (2021) *Handbook for Quality and Metadata Report, 2021 re-edition.* Publications Office of the European Union, Luxembourg

Freedman D., Diaconis P. (1981) "On the histogram as a density estimator: L2 theory". *Probability Theory and Related Fields,* 57, pp. 453–476

Korn E.I., Graubard B.I. (1999) *Analysis of Health Surveys*. Wiley, New York.

O'Muircheataigh C., Pedlow S. (2002) "Combining samples vs. cumulating cases: a comparison of two weighting strategies in NLS97". *American Statistical Association Proceedings of the Joint Statistical Meetings*, pp. 2557-2562.

Sarndal C.E., Swensson B., Wretman J.H. (1992) *Model Assisted Survey Sampling*. Springer–Verlag, New York.

Silverman B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman & Hall

Istat