Contrafactual Degrees of Freedom

Phillip S. Kott
pkott@rti.org

$$F_{contrafactual} = \frac{\left(\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{k\in s_{hj}} w_k^2\right)^2}{\sum_{h=1}^{H}\left[\sum_{j=1}^{n_h}\left(\sum_{k\in s_{hj}} w_k^2\right)^2 + \frac{1}{(n_h-1)^2}\left\{\left[\sum_{j=1}^{n_h}\left(\sum_{k\in s_{hj}} w_k^2\right)\right]^2 - \sum_{j=1}^{n_h}\left(\sum_{k\in s_{hj}} w_k^2\right)^2\right\}\right]}$$

Introduction

The nominal degrees of freedom of a variance estimator under design-based survey sampling theory

The Contrafactual degrees of freedom

called the *effective degrees of freedom* in

Kott (*Survey Methodology*, 1994) and

Korn and Graubard (*Analysis of Health Surveys*, 1999)

Variants

A regression coefficient

An empirical example

Some comments

Background

When constructing coverage (confidence) intervals for an estimated mean or regression coefficient (m), design-based practice is to use a *stratified sandwich variance estimator* ($\hat{v}$) with (nominal) degrees of freedom equal to

The number of primary sampling units minus the number of strata (*minus the number of regressors plus 1*).

Replication produces the same value.

This practice has little theoretical justification except under a very unlikely *outcome* model.

Background (2)

From a model-based point of view, the stratified sandwich variance estimator allows

Nonnormal and heteroscedastic errors, and

almost arbitrary clustering within PSUs.

When estimating means, using nominal degrees of freedom implicitly assumes:

PSU-level aggregates are normal and identically distributed, and

Strata are nuisances.

The Contrafactual Degrees of Freedom

assumes (contra factually) that the elements are normal and *iid,* and that the relative variance of $\hat{v}$ is 2/its degrees of freedom (as in a chi-squared distribution – a Satterthwaite approximation).

So

$$F_{contrafactual} = \frac{\left(\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{k\in s_{hj}} w_k^2\right)^2}{\sum_{h=1}^{H}\left[\sum_{j=1}^{n_h}\left(\sum_{k\in s_{hj}} w_k^2\right)^2 + \frac{1}{(n_h-1)^2}\left\{\left[\sum_{j=1}^{n_h}\left(\sum_{k\in s_{hj}} w_k^2\right)\right]^2 - \sum_{j=1}^{n_h}\left(\sum_{k\in s_{hj}} w_k^2\right)^2\right\}\right]},$$

where w_k is the weight of element k, and

s_{hj} is the (respondent) sample in PSU j in stratum h (of H).

There are n_h PSUs in stratum h.

Variants

The weight can be a design weight or a calibrated weight.

Each weight can be multiplied by a domain-inclusion indicator for a domain mean.

Element-level weights can be replaced by PSU-level sums of weights assuming complete correlation within each PSU.

Each weight can be multiplied by an element-level standard error.

Alternative Formula and Another

Let $w_{hj}^{(2)} = \sum_{k \in s_{hj}} w_k^2$, then

$$F_{cont.} = \frac{\left(\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj}^{(2)}\right)^2}{\sum_{h=1}^{H} \left[\sum_{j=1}^{n_h} w_{hj}^{(2)^2} + \frac{1}{(n_h - 1)^2} \left\{\left[\sum_{j=1}^{n_h} w_{hj}^{(2)}\right]^2 - \sum_{j=1}^{n_h} w_{hj}^{(2)^2}\right\}\right]}$$

Contrafactual Degrees of Freedom for a Domain Mean treating the $w_{hj}^{(2)}$ *as 0 or 1* (i.e., the PSU is empty or not empty)

number of nonempty PSUs in h

$$F_{contrafactual} = \frac{\sum_{h=1}^{H} d_h}{\left(1 + \frac{\sum_{h=1}^{H} d_h (d_h - 1)/(n_h - 1)^2}{\sum_{h=1}^{H} d_h}\right)}$$

A Regression Coefficient

b_2 in $\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ b_2 \end{pmatrix}$ the solution to

$$\sum_{k \in S} w_k \mathbf{z}_k \left[y_k - f(\mathbf{z}_k{}^T \mathbf{b}) \right] = \mathbf{0}$$

Define $z_{2k}{}^* = z_{2k} - \mathbf{z}_{1k} (\sum_S w_j f_j' \, \mathbf{z}_{1j} \, \mathbf{z}_{1j}{}^T)^{-1} \sum_S w_j f_j' \mathbf{z}_{1j} \, z_{2j}$.

Replace w_k with $w_k \, z_{2k}{}^*$.

The difference between domain means can be estimated as a regression coefficient.

An Empirical Example

In Kott (2020), I reorganized the MU281 population of 281 Swedish administrative municipalities (from the big Yellow Book) into 20 strata containing 108 clusters.

Using simple random sampling *with* replacement,
I drew 2500 fresh simulated samples containing four clusters per stratum.

I looked at estimated variances ($\hat{v}$) for estimates (denoted $\hat{m}$) among municipalities where at least 50% of the municipal council in 1982 were Social Democrats, that is, Group 1:

A. The mean population size in 1985.

B. The fraction of municipalities in 1985 with tax/person rates at or above 8,000 Kronor.

Results for the Group-1 Means

My computations of the contrafactual degrees of freedom are (implied *t*-value at two-sided 95% confidence level)

	Elements independent	PSUs independent
Mean	20.6 (2.08)	14.2 (2.14)
Min	14.4 (2.14)	8.1 (2.30)
Max	27.3 (2.05)	20.2 (2.08)

Nominal: 60 (2.00)

Empirical F = 2/emp. relVar($\hat{v}$): A = 27.4 (2.10), B = 8.2 (2.37)

(where the cumulative distribution of $\left|\frac{\hat{m}-m}{\sqrt{\hat{v}}}\right|$ is at 95%)

Results for Differences in Group Means

My computations of the contrafactual degrees of freedom are (implied *t*-value at two-sided 95% confidence level)

	Elements independent	PSUs independent
Mean	31.0 (2.04)	16.0 (2.14)
Min	24.3 (2.06)	8.3 (2.29)
Max	36.5 (2.03)	23.3 (2.07)

Nominal: 59 (2.00)

Empirical F = 2/emp. relVar($\hat{v}$): A = 34.0 (2.20), B = 14.5 (2.21)

(where the cumulative distribution of $\left|\frac{\hat{m}-m}{\sqrt{\hat{v}}}\right|$ is at 95%)

Some Comments

Although better than the nominal, the contrafactual degrees of freedom is not perfect (hence not "effective") because:

The variance estimators aren't chi-squared.

The estimators and their estimated variances are asymptotically unbiased, but the world is finite.

The data may not be normally distributed.

Weights may not be ignorable.

Strata may not be ignorable.

Within-cluster correlations may be complex.

Selection of clusters was *with* replacement.

Some Comments

When estimating a regression model, it may be reasonable to assume a model.

In particular a model, where

Strata are ignorable (which is easier when there are a lot more PSUs than strata).

Within-cluster correlations only occur at a level below the PSU (e.g., within households or blocks as opposed to counties).

When testing groups of coefficients, consider using a series of Bonferroni-adjusted *t* tests in place of an adjusted Wald F test.

References

Korn E. and Graubard, B. (1999). *Analysis of Health Surveys*. New York: John Wiley (especially Chapter 5)

Kott, P. (2020). *The degrees of freedom of a variance estimator in a probability sample.* Research Triangle Park, NC: RTI Press

Kott, P. (1994). Hypothesis testing of linear regression coefficients with survey data. *Survey Methodology*, 159–164