Applying chi-square tests to examine homogeneity of proportions between data collected with different sampling designs- a simulation study

Li-Yen Rebecca Hu, Van L. Parsons, Yulei He, Katherine E. Irimata, Rong Wei

Division of Research and Methodology, National Center for Health Statistics,

U.S. Centers for Disease Control and Prevention

October 24, 2024

Study question:

- Chi-square tests are often employed to examine
  - Association of categorical variables
  - Homogeneity of proportions
  - Goodness-of-fit for a specified distribution
- Previously we examined programming codes and concepts of four types of chi-square tests with four modern statistical packages applicable to all three purposes. (The manuscript is under review by The American Statistician.)
- This presentation focuses on the actual performance (i.e., Type 1 and Type 2 errors) in simulations of two types of chi-square tests in examining <u>homogeneity of proportions</u> acquired by two samples with different sampling designs.

Types of chi-square tests or their related F tests examined using the *survey* package in R (4.4-2)

- Wald test
  - Generalized F test
  - Adjusted F test
- Rao-Scott (RS) chi-square test
  - First-order chi-square test
  - Second-order F test

Pearson chi-square test:

$$Q_P = \left(\frac{n}{\widehat{N}}\right) \sum_r \sum_c \frac{\left(\widehat{N}_{rc} - E_{rc}\right)^2}{E_{rc}}$$

n – total sample size
r – row
c – column
$\widehat{N}$, $\widehat{N}rc$ -estimated weighted overall total and cell frequency
E_{rc} – expected weighted cell frequency

Rao-Scott chi-square test: 1st order chi-square test

$$Q_{RS1} = \frac{Q_P}{D}$$

$$D = \frac{\sum_r \sum_c (1 - \hat{p}_{rc}) d_{rc} - \sum_r (1 - \hat{p}_{r\cdot}) d_{r\cdot} - \sum_c (1 - \hat{p}_{\cdot c}) d_{\cdot c}}{(R-1)(C-1)}$$

$$d_{rc} = \widehat{Var}(\hat{p}_{rc}) / \widehat{Var}_{\mathrm{SRS}}(\hat{p}_{rc})$$

Degrees of freedom (DF): (R-1)(C-1)

R - number of rows; C- number of columns

$\hat{p}_{r\cdot}$, $\hat{p}_{\cdot c}$ - marginal probability estimates for row *r* and column *c*, respectively

$d_{r\cdot}$, $d_{\cdot c}$ - design effects for row *r* and column *c*, respectively

Rao-Scott chi-square test: 2nd order F-test

$$F_{RS2} = \frac{Q_{RS1}}{(R-1)(C-1)}$$

DF: Numerator: $\frac{(R-1)(C-1)}{1+\hat{a}^2}$; Denominator: $\frac{s(R-1)(C-1)}{1+\hat{a}^2}$

$\hat{a}^2 = \left(\sum_{i=1}^{K} \frac{d_i^2}{K\bar{d}^2}\right) - 1$; s – DF for the variance estimator

i – individual cells of the contingency table

$K = (R-1)(C-1)$; $\bar{d}$- the average eigenvalue

d_i- eigenvalues of the estimated design effects matrix

Wald test

$$Q_W = \hat{\mathbf{Y}}^{\mathrm{T}}\left[\hat{\mathbf{V}}\left(\hat{\mathbf{Y}}\right)\right]^{-1}\hat{\mathbf{Y}}$$

$$F_W = \frac{Q_W}{(R-1)(C-1)}$$

DF: Numerator: $(R - 1)(C - 1)$;

Denominator: DF for the variance estimator

$\hat{\boldsymbol{Y}}$- an $(R-1)(C-1)$ array of $\hat{Y}_{rc}$

$\hat{Y}_{rc} = \hat{N}_{rc} - E_{rc}$

$\hat{\boldsymbol{V}}(\hat{\boldsymbol{Y}})$- the design-consistent variance-covariance matrix for $\hat{\boldsymbol{Y}}$

Adjusted Wald test

$$F_{adjW} = \frac{Q_W(s-k+1)}{ks}$$

DF: Numerator: $(R - 1)(C - 1)$

Denominator: $s - k + 1$

$k = (R - 1)(C - 1)$

s – DF for the variance estimator

Research Components:

1. The performance of the variants of the Wald test and the RS chi-square test with three combinations of three sampling designs and with six variations of a 5-category outcome variable.

2. The power of the variants of the Wald test and the RS chi-square test with three combinations of two sampling designs.

3. Weight adjustment strategies when combining two samples with different sampling designs.

Population Generated

- Total population units: 100,000
- Five equally-sized strata with 200 clusters in each
- Five types of clusters with different measure of size (MOS)
  - M1 (10 clusters): 300 units
  - M2 (20 clusters): 200 units
  - M3 (60 clusters): 100 units
  - M4 (60 clusters): 75 units
  - M5 (50 clusters): 50 units
- 6 variations of outcomes : y1A, y1B (2 versions), y1C (2 versions), y1E

Reiter et al. (2006) Surv. Methodol. 32:143-149.

Population Generated (continued)

- Targeting overall multinomial probability:

  p0 = (0.15, 0.20, 0.15, 0.25, 0.25)
- y1A- independent of strata, clusters, MOS
- y1B- independent of clusters, but MOS-dependent
  - independent of strata (Set 1 population)
  - strata-dependent (Set 2 population)
- y1C- cluster- and MOS-dependent
  - independent of strata (Set 1 population)
  - strata-dependent (Set 2 population)
- y1E- independent of strata, clusters, MOS

  p = (0.20, 0.20, 0.20, 0.20, 0.20)

Part 1: Sampling Design- 2000 draws

- Simple random sampling (SRS) – 1,000 sampled units
- Complex sampling design 1 (CSD1) – 1,000 sampled units
  - varied numbers of primary sampling units (PSU) and secondary sampling units (SSU) per strata
  - roughly same total numbers of units selected per stratum
  - equal probability of selection method (EPSEM)
- Complex sampling design 2 (CSD2) – 1,000 sampled units
  - varied numbers of PSU and SSU
  - different numbers of units selected per stratum (non-EPSEM)
- Samples combined as three pairs: CSD1 vs SRS; CSD2 vs SRS; CSD1 vs CSD2
- Combined data have equal sums of weights contributed from each sample, with the sum of the final adjusted weights equals to the total number of population units

Part 1 Result Summary

- The differences in results acquired by the four variants of tests are *minor*.
- The second-order RS F test is generally the most conservative (i.e, more likely to attain larger p-values) among the four
- The Wald test and the adjusted Wald test are slightly more liberal (i.e, more likely to attain smaller p-values)

Part 1 Result Summary (continued)

- For a variable with categories of equal probabilities, type 1 error rates of all variants of tests examined are slightly *inflated*.
  - Conservative tests like the 2nd-order RS F test show performance slightly closer to the specified 5% Type 1 error rate.
- For a variable with categories of unequal probabilities, especially when both samples are of complex sampling design, all four variants of tests examined tend to exhibit *lower* Type 1 error rate than the specified 5%.
  - The Wald test and the adjusted Wald test show performance closer to the specified 5% Type 1 error rate than the other two.

Part 2: Power Analysis-Population

- Using the Set 2 Population as in Part 1
- To identify which method performs better to detect the differences in underlying multinomial probabilities.

$\vec{a}$	$\overrightarrow{p0}$ or $\overrightarrow{p0^*}$	$\overrightarrow{p0E}$ or $\overrightarrow{p0E^*}$
-	(0.150, 0.200, 0.150, 0.250, 0.250)	(0.200, 0.200, 0.200, 0.200, 0.200)
(3.5, 3.0, 3.5, 3.2, 3.0)	(0.164, 0.188, 0.164, 0.250, 0.234)	(0.216, 0.185, 0.216, 0.198, 0.185)
(3.8, 3.3, 4.3, 3.8, 5.0)	(0.140, 0.162, 0.158, 0.233, 0.307)	(0.188, 0.163, 0.213, 0.188, 0.248)

$$\overrightarrow{p0^*} = \frac{\vec{a} * \overrightarrow{p0}}{\sum(\vec{a} * \overrightarrow{p0})}$$

Part 2: Power Analysis-Sampling (2000 draws)

- Samples selected with the following designs:
  - SRS (p0) vs SRS (p0_A, p0_B)
  - CSD2 (p0) vs SRS (p0_A, p0_B)
  - CSD2 (p0) vs CSD2 (p0_A, p0_B)
- Each sample has 1000 units
- Combined data have equal sums of weights contributed from each sample, with the sum of the final adjusted weights equals to the total number of population units

Part 2 Result Summary

- The Wald and the adjusted Wald tests perform slightly better at detecting different underlying multinomial distributions.
- In most scenarios examined, when the sample sizes are fixed, SRS samples have better power in detecting different underlying multinomial distributions, especially for variables that are strata- and cluster-dependent.

Part 3: Weight adjustment

Set 2 Population – Same as in Part 1

Sampling Design - 2000 draws

- Simple random sampling (SRS) – 200 sampled units
- Complex sampling design 1 (CSD1) – 1,800 sampled units
  - EPSEM
- Complex sampling design 2 (CSD2) – 1,800 sampled units
  - non-EPSEM
- Samples combined as two pairs: CSD1 vs SRS; CSD2 vs SRS

Part 3: Weight adjustment

- Wt_{orig} - original weight from each sample, unadjusted
- Wt_{50} – equal sums of weights (50:50) from each sample, with the sum of the final adjusted weights equal to the total number of population units
- Wt_{eff} - effective sample size adjustment; the ratio of sum of weights from each sample is reflective of their effective sample sizes; effective sample sizes are computed as nominal sample sizes divided by design effect

$$design\ effect = 1 + \left(\frac{standard\ deviation\ of\ Wt_{orig}}{mean\ Wt_{orig}}\right)^2$$

- Wt_{nom} - nominal sample size adjustment; the ratio of sum of weights from each sample is reflective of their nominal sample sizes

Part 3 Result Summary

- In our simulation setting, no difference was observed between Wtorig and Wt50
- Weights adjusted with effective and nominal sample sizes could help the 1st order and 2nd order RS chi-square tests getting closer to the specified 5% Type 1 error rate.
- For a variable independent of strata and clusters, the 1st order and 2nd-order RS chi-square tests show performance close to the specified 5% Type 1 error rate.
- For a variable dependent on strata and/or clusters, the Wald and the adjusted Wald tests show performance close to the specified 5% Type 1 error rate.

Overall Summary

- In our simulation settings examined for testing homogeneous proportions:
- The differences in results acquired by the four tests are minor
- The 2nd order RS F test is generally the most conservative among the four
- The Wald test and the adjusted Wald test are slightly more liberal than the Rao-Scott adjusted tests.

Overall Summary (continued)

- If a variable is dependent on strata and/or clusters, the Wald and the adjusted Wald tests show performance close to the specified 5% Type 1 error rate
- If a variable is equally distributed between each category and is independent of strata and/or clusters, the 1^{st} and 2^{nd} order RS tests exhibit performance close to the specified 5% Type 1 error rate.
- The Wald and the adjusted Wald tests perform slightly better at detecting different underlying multinomial distributions.

Future Directions

- Evaluate the performance of different types of chi-square or F tests with different multinomial probabilities.
- Examine the performance of other types of chi-square or related F tests, such as the Rao-Scott Likelihood Ratio Chi-Square Test and the Wald Log-Linear Chi-Square Test that are not available in the R *survey* package.
- Examine the performance of each type of chi-square or related F tests with real data

Thank you!

Questions or comments? Please e-mail liyenhu@cdc.gov

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

Appendix: Numerical Results

1. The performance of the variants of Wald test and RS chi-square test with three combinations of three sampling designs and with six variations of a 5-category outcome variable.
2. The power of the variants of the Wald test and the RS chi-square test with three combinations of two sampling designs.
3. Weight adjustment strategies when combining two samples with different sampling designs.

Interpretation of results

The results listed in the remaining slides are computed as the percentage of number of simulations out of 2000 simulation runs reported p-values < 0.05. If a method performs well, the percentage value should be ~ 0.05.

Part 1: Simulation Results of y1A

Sample Pair†	Pop	1st order RS $\chi 2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
1S	1	5%	5%	5%	5%
2S	1	5%	5%	5%	5%
12	1	5%	5%	5%	5%
1S	2	5%	5%	6%	6%
2S	2	4%	4%	5%	5%
12	2	4%	4%	5%	5%

†1S - CSD1 vs SRS; 2S - CSD2 vs SRS; 12 - CSD1 vs CSD2

Part 1: Simulation Results of y1B

Sample Pair†	Pop	1st order RS $\chi 2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
1S	1	4%	3%	4%	4%
2S	1	4%	4%	5%	5%
12	1	3%	3%	3%	3%
1S	2	3%	3%	4%	4%
2S	2	4%	4%	5%	5%
12	2	2%	2%	3%	3%

†1S - CSD1 vs SRS; 2S - CSD2 vs SRS; 12 - CSD1 vs CSD2

Part 1: Simulation Results of y1C

Sample Pair†	Pop	1st order RS $\chi2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
1S	1	5%	4%	5%	5%
2S	1	5%	5%	6%	6%
12	1	4%	4%	5%	4%
1S	2	4%	4%	5%	5%
2S	2	4%	4%	4%	4%
12	2	4%	3%	5%	4%

†1S - CSD1 vs SRS; 2S - CSD2 vs SRS; 12 - CSD1 vs CSD2

Part 1: Simulation Results of y1E

Sample Pair†	1st order RS $\chi 2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
1S	6%	6%	6%	6%
2S	5%	5%	6%	6%
12	6%	6%	6%	6%

†1S - CSD1 vs SRS; 2S - CSD2 vs SRS; 12 - CSD1 vs CSD2

Appendix: Numerical Results

1. The performance of the variants of the Wald test and the RS chi-square test with three combinations of three sampling designs and with six variations of a 5-category outcome variable.

2. The power of the variants of the Wald test and RS chi-square test with three combinations of two sampling designs.

3. Weight adjustment strategies when combining two samples with different sampling designs.

Part 2: Simulation Results of y1A

p0 vs	Sample Pair	1st order RS $\chi 2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
p0_A	SRS, SRS	19%	19%	20%	20%
	CSD2, SRS	19%	19%	20%	20%
	CSD2, CSD2	20%	20%	21%	21%
p0_B	SRS, SRS	67%	67%	67%	67%
	CSD2, SRS	63%	62%	64%	64%
	CSD2, CSD2	61%	60%	62%	61%

Part 2: Simulation Results of y1B

p0 vs	Sample Pair	1st order RS $\chi 2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
p0_A	SRS, SRS	19%	19%	19%	19%
	CSD2, SRS	15%	14%	17%	17%
	CSD2, CSD2	11%	10%	13%	12%
p0_B	SRS, SRS	75%	75%	75%	75%
	CSD2, SRS	68%	67%	66%	66%
	CSD2, CSD2	56%	54%	56%	55%

Part 2: Simulation Results of y1C

p0 vs	Sample Pair	1st order RS $\chi 2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
p0_A	SRS, SRS	19%	19%	19%	19%
	CSD2, SRS	15%	15%	18%	17%
	CSD2, CSD2	11%	11%	14%	14%
p0_B	SRS, SRS	76%	76%	77%	77%
	CSD2, SRS	68%	67%	68%	68%
	CSD2, CSD2	57%	56%	58%	57%

Part 2: Simulation Results of y1E

p0 vs	Sample Pair	1st order RS $\chi 2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
p0E_A	SRS, SRS	20%	20%	20%	20%
	CSD2, SRS	20%	19%	20%	20%
	CSD2, CSD2	17%	17%	19%	18%
p0E_B	SRS, SRS	63%	63%	64%	63%
	CSD2, SRS	58%	58%	59%	59%
	CSD2, CSD2	55%	54%	56%	56%

Appendix: Numerical Results

1. The performance of the variants of the Wald test and the RS chi-square test with three combinations of three sampling designs and with six variations of a 5-category outcome variable.

2. The power of the Wald test and the RS chi-square test with three combinations of two sampling designs.

3. Weight adjustment strategies when combining two samples with different sampling designs.

Part 3: Simulation Results of y1A

Sample	Weight	1st order RS $\chi 2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
CSD1 vs SRS	Wtorig	5%	5%	6%	6%
	Wt50	5%	5%	6%	6%
	Wteff	5%	5%	6%	6%
	Wtnom	5%	5%	6%	6%
CSD2 vs SRS	Wtorig	5%	5%	6%	5%
	Wt50	5%	5%	6%	5%
	Wteff	5%	5%	6%	5%
	Wtnom	5%	5%	6%	5%

Part 3: Simulation Results of y1B

Sample	Weight	1st order RS $\chi2$ Test	2nd order RS F test	Wald Test	Adjusted Wald Test
CSD1 vs SRS	Wtorig	4%	4%	5%	5%
	Wt50	4%	4%	5%	5%
	Wteff	5%	4%	5%	5%
	Wtnom	5%	4%	5%	5%
CSD2 vs SRS	Wtorig	4%	4%	5%	5%
	Wt50	4%	4%	5%	5%
	Wteff	4%	4%	5%	5%
	Wtnom	4%	4%	5%	5%

Part 3: Simulation Results of y1C

Sample	Weight	1^{st} order RS $\chi 2$ Test	2^{nd} order RS F test	Wald Test	Adjusted Wald Test
CSD1 vs SRS	Wtorig	4%	4%	5%	5%
	Wt50	4%	4%	5%	5%
	Wteff	5%	4%	5%	5%
	Wtnom	5%	4%	5%	5%
CSD2 vs SRS	Wtorig	3%	3%	4%	4%
	Wt50	3%	3%	4%	4%
	Wteff	4%	4%	4%	4%
	Wtnom	4%	4%	4%	4%

Part 3: Simulation Results of y1E

Sample	Weight	1^{st} order RS $\chi 2$ Test	2^{nd} order RS F test	Wald Test	Adjusted Wald Test
CSD1 vs SRS	Wtorig	5%	5%	6%	6%
	Wt50	5%	5%	6%	6%
	Wteff	5%	5%	6%	6%
	Wtnom	5%	5%	6%	6%
CSD2 vs SRS	Wtorig	5%	4%	6%	6%
	Wt50	5%	4%	6%	6%
	Wteff	5%	5%	6%	6%
	Wtnom	5%	5%	6%	6%