



Re-scaling weights to Effective Sample Size

PHUSE EU Connect 2020
9th–13th November

Agenda

- Introduction & questions
- Definitions
 - ✓ Sample Size
 - ✓ Weighting methods
 - ✓ Effective Sample Size
 - ✓ Design Effect
- Answers
 - ✓ Example of explanation of the problem
 - ✓ Demonstration
- Conclusion
- References

Introduction

What is the question we start from?

We have two:

How to calculate the sample size for a weighted study?

How to create proper inference on the weighted data?

Definitions: Weighting methods, why use them

Similarity

For indirect treatment comparison a similarity assumption is required in terms of moderators of treatment effect (i.e. patients included should be sufficiently similar in the two sets of treatments compared)

How is possible to reach similarity?



Uniforming:

- Inclusion/exclusion criteria
- Outcome definitions
- **Weighting baseline characteristics**
- Statistical methods



Definitions: Weighting methods

The propensity score is calculated using logistic regression by modeling the probability of enrolment to TRIAL-1.

Covariates will include all characteristics that could affect the probability of enrolment to a study but not the outcome of interest. The probability for each person from the logistic regression model is considered the propensity to be enrolled to the study. The estimated weights incorporate the propensity score to weight each group to an analyst-selected target population.

Table 1. Examples of balancing weights and corresponding target population and estimand under different h .

Target population	$h(x)$	Estimand	Weight (w_1, w_0)
Combined	1	ATE	$\left(\frac{1}{e(x)}, \frac{1}{1-e(x)}\right)$ [HT]
Treated	$e(x)$	ATT	$\left(1, \frac{e(x)}{1-e(x)}\right)$
Control	$1 - e(x)$	ATC	$\left(\frac{1-e(x)}{e(x)}, 1\right)$
Overlap	$e(x)(1 - e(x))$	ATO	$(1 - e(x), e(x))$
Truncated combined	$1(\alpha < e(x) < 1 - \alpha)$		$\left(\frac{1(\alpha < e(x) < 1 - \alpha)}{e(x)}, \frac{1(\alpha < e(x) < 1 - \alpha)}{1 - e(x)}\right)$
Matching	$\min\{e(x), 1 - e(x)\}$		$\left(\frac{\min\{e(x), 1 - e(x)\}}{e(x)}, \frac{\min\{e(x), 1 - e(x)\}}{1 - e(x)}\right)$

where $h(\cdot)$ is prespecified function of x

Andrea Parodi, Dmitri Petratchenko, Valos srl, Genova.

Re-scaling weights to Effective Sample Size

Li et al. (2018)

Definitions: Weighting methods, rescaled weights

The estimated weights may be further rescaled so that the rescaled weights are relative to the original unit weights of each individual (a rescaled weight >1 means that an individual carries more weight in the reweighted population than in the original population, and a rescaled weight <1 means that an individual carries less weight). Given sample size n estimated weights, the rescaled weight for individual j is calculated as

$$w.rs_j = n * \frac{w_j}{\sum_{i=1}^n w_j} .$$

Definitions: Effective Sample Size

The effective sample size (ESS) for the weighed population data suggested by [Signorovitch 2010] will be calculated using the formula

$$ESS = \frac{(\sum_{i=1}^n w_i)^2}{\sum_{i=1}^n w_i^2}$$

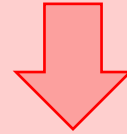
Given sample size n and w the estimated weights with $i=1, \dots, n$

Although the effective sample size ESS is only accurate if the weights are fixed (or not related to outcomes), smaller sample sizes are an indication that the weights are highly variable due to a lack of population overlap and the estimates may be unstable.

Definitions: Sample Size

To determine a sample size, some criterion must be adopted for deciding how big is big enough. This is a question of how precise you want an estimate to be.

Complex sample designs



These designs have features such as multistage sampling, selection with unequal probabilities, sample weighting adjustments and population weighting adjustments, this causes subgroups represented disproportionately in the sample.



Design effect concept



Effective Sample Size

Definitions: Design Effect (DE)

The concept of design effect (DE) has proved to be a valuable tool in the design of complex samples and helps guide the development of efficient sample designs.

$$DE = \frac{\text{Actual Sample Size}}{\text{Simple Random Sample}}$$

It was proposed [Cornfield, 1951] that sample size for complex surveys can be calculated by first calculating the sample size required for a Simple Random Sample (SRS) and then adjusted by the design effect, e.g. if the DE is equal to 2, this means that sample size be at least doubled.

Question

How to calculate the sample size for a weighted study and how to create proper inference on the weighted data?

In this presentation, 4 cases are compared: unweighted data, weighted data, weighted data with re-scaled weights where sum of weights is equal to the number of subjects in the sample and weighted data with re-scaled weights where sum of weights is equal to ESS. Each case is implemented on the same simulated dataset.

The purpose of this work is to demonstrate why it is better to re-scale weights using ESS and how the sample size should be calculated to have reliable results. The comparisons are conducted on a simulated example estimating risk ratio as result.

Example of explanation of the problem

Let's suppose to have two RWD cohorts, the objective is to assess if the percentage of responders is bigger in cohort A taking drug A than in cohort B taking drug B.

The results are:

	Responders	Non-responders	Total
Cohort A	40	36	76
Cohort B	43	97	140

Calculating risk ratio with log-binomial model we get:

$RR(95\%) = 1.7136 (1.2348, 2.3781)$ $p\text{-value} = 0.0013$

Example of explanation of the problem

Calculating risk ratio with log-binomial model weighted with overlap weights we get:

	Responders	Non-responders	Total
Cohort A	20.7152	18.9842	39.6993
Cohort B	12.2569	27.4424	39.6993

$RR(95\%) = 1.6901 (0.9726, 2.9368)$ $p\text{-value} = 0.0627$

Example of explanation of the problem

Calculating risk ratio with log-binomial model weighted with re-scaled (to original 216 subjects) overlap weights we get:

	Responders	Non-responders	Total
Cohort A	56.3546	51.6454	108
Cohort B	33.3444	74.6556	108

$RR(95\%) = 1.6901 (1.2090, 2.3627)$ $p\text{-value} = 0.0021$

What is the right result?

Example of explanation of the problem

The solution could be re-scaling not to 216 but to Effective Sample Size, keeping the proportions of subjects in cohort A and B the same of original ones.

The ESS is 159.72226767

	Responders	Non-responders	Total
Cohort A	29.3245	26.8741	56.1986
Cohort B	31.9623	71.5613	103.524

RR(95%) = 1.6901 (1.1538, 2.4757)

p-value = 0.0071

Demonstration: step 1

Simulate a population of 10'000'000 subjects based the following assumption:
proportion of successes in treatment 1 = 20%, in treatment 2 = 30%.

Table of treatment by success			
treatment	success		
	0	1	Total
1	4000376 80.01	999624 19.99	5000000
2	3500493 70.01	1499507 29.99	5000000
Total	7500869	2499131	1E7

Demonstration: step 2

Based on the book *Wang, H. and Chow, S.-C. 2007. Sample Size Calculation for Comparing Proportions. Wiley Encyclopedia of Clinical Trials*, $n_1 = n_2 = 291$ subjects are needed for showing a difference in the two groups if $p_1 = 0.2$ and $p_2 = 0.30$ with $\alpha = 0.05$ and power = 0.8.

So we randomly extract 10'000 samples, expecting to find a significant difference in 80% of them.

result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H0 not rejected	2147	21.47	2147	21.47
H0 rejected	7853	78.53	10000	100.00

Demonstration: step 2

We tried increasing $n1 = n2$ starting from 292 until we got the first value with $>80\%$ of successes.

We found for $n1 = n2 = 296$

result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H0 not rejected	2088	20.88	2088	20.88
H0 rejected	7912	79.12	10000	100.00

And for $n1 = n2 = 297$

result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H0 not rejected	1971	19.71	1971	19.71
H0 rejected	8029	80.29	10000	100.00

So 297 seems to be the correct sample size needed.

Demonstration: step 3

Create a population with $p1 = 0.2$ and $p2 = 0.3$, but with additional effect of age, sex and severity class. Results of the population created are:

Table of treatment by success			
treatment	success		
	0	1	Total
1	3892261 77.85	1107739 22.15	5000000
2	3472547 69.45	1527453 30.55	5000000
Total	7364808	2635192	1E7

Demonstration: step 3

- Age is created as uniform distribution with range 40-80 in group 1, 45-85 in group 2.
- Female cat is created as bernoulli distribution with $p = 0.6$ in group 1, $p = 0.5$ in group 2.
- Severity class severe is created as bernoulli distribution with $p = 0.3$ in group 1, $p = 0.4$ in group 2.
- Starting from 61st year the probability of experiencing a success has been decreased of 0.001 for each year (increased for ages < 60).
- Be in the female group increased the probability of experiencing a success of 0.06.
- Be in the severe group decreased the probability of experiencing a success of 0.05.

Table of treatment by Female			
treatment	Female		
	0	1	Total
1	2000935 40.02	2999065 59.98	5000000
2	2499828 50.00	2500172 50.00	5000000
Total	4500763	5499237	1E7

Table of treatment by severity			
treatment	severity		
	0	1	Total
1	3500295 70.01	1499705 29.99	5000000
2	2999965 60.00	2000035 40.00	5000000
Total	6500260	3499740	1E7

Demonstration: step 4

Find $n1 = n2$ that has mean/median ESS > 594 (297×2).

I found $n1 = n2 = 318$, with mean ESS = 592.53, median ESS = 593.28.

The following group $n1 = n2 = 319$ had mean = 594.18 and median = 594.86.

So design effect can be estimated as $319/297 = 1.074$

Demonstration: step 5

Simulate the number of studies with success using weights in the dataset with $n1 = n2 = 319$ created at step 4.

Without using weights:

result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H0 not rejected	3141	31.41	3141	31.41
H0 rejected	6859	68.59	10000	100.00

Using weights:

result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H0 not rejected	4893	48.93	4893	48.93
H0 rejected	5107	51.07	10000	100.00

Demonstration: step 5

Using re-scaled to $n1 + n2$ weights

result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H0 not rejected	1823	18.23	1823	18.23
H0 rejected	8177	81.77	10000	100.00

Using re-scaled to ESS weights:

result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H0 not rejected	2026	20.26	2026	20.26
H0 rejected	7974	79.74	10000	100.00

The closest to 80.29 is the ess-scaling to $n1 + n2$ weights

Conclusion

- Rescaling ESS seems to be the most reliable solution even if maybe the evidence found in this simulation is not strong enough
- Evidence is strong enough for stating that weights without rescaling should not be used
- Currently a mathematical demonstration is lacking
- Need of evidence using different weighting methods and for comparisons with only 1 study with IPD
- In the proposed solution the ESS of ~ 160 subjects has been divided in Cohort A and B keeping the proportion equal to the original one without weighting. Further simulations are needed for checking different situations, for example checking ESS calculated singularly on each of the two cohorts

Thank you

REFERENCES

- Kish, L.: 1965, Survey Sampling, John Wiley & Sons.
- Cornfield J. The determination of sample size. Am J Public Health Nations Health. 1951;41(6):654–61.
- Signorovitch et al. (2010). Comparative Effectiveness Without Head-to-Head Trials. Pharmacoeconomics, 28(10), 935.
- Fujian Song (2009). What is indirect comparison?. Hayward Medical Communications, a division of Hayward Group Ltd.
- Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing Covariates via Propensity Score Weighting. Journal Of The American Statistical Association, 2018, VOL. 113, NO. 521, 390-400, Theory and Methods.
- Wang, H. and Chow, S.-C. 2007. Sample Size Calculation for Comparing Proportions. Wiley Encyclopedia of Clinical Trials