General Psychiatry

Inconsistency between overall and subgroup analyses

Hongyue Wang,[1] Bokai Wang ,[1] Xin M Tu,[2] Changyong Feng [1]

To cite: Wang H, Wang B, Tu XM, *et al*. Inconsistency between overall and subgroup analyses. *General Psychiatry* 2022;**35**:e100732. doi:10.1136/gpsych-2021-100732

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/gpsych-2021-100732).

Received 21 February 2022
Accepted 27 April 2022

[1]Departments of Biostatistics & Computational Biology, University of Rochester, Rochester, New York, USA
[2]Division of Biostatistics and Bioinformatics, Herbert Wertheim School of Public Health and Human Longevity Science, UC San Diego, La Jolla, California, USA

Correspondence to
Professor Changyong Feng; Changyong_Feng@URMC.Rochester.edu

ABSTRACT

Suppose we have a sample of subjects in two treatment groups. To study the difference of the treatment effects, we can analyse the data using all subjects (overall analysis). We may also divide the subjects into several subgroups based on some covariates of interest (eg, gender), and study the treatment effects within each subgroup. The results of these two analyses may be different or even in opposite directions. In this paper, we give a general sufficient condition of consistency between the overall and subgroup analyses.

INTRODUCTION

Consider the following hypothetical example. Suppose the fourth grade students of two schools (1 and 2) in a school district took a state maths exam. The principals of these two schools wanted to know whether there was a difference in the performances between the two schools. They calculated the overall average score, the average score of girls and the average score of boys in each school. The result is presented as scenario B in table 1. After looking at the average scores of girls and boys, respectively, the principal of school 1 was very happy as they were both one point higher than those in school 2. However, after looking at the overall average score of these two schools, the principal was very confused as the overall average score of school 2 was higher than that of school 1. Is there anything wrong in calculating the overall average score? What is the reason for the inconsistency between the overall average scores and the average scores stratified by gender?

Before figuring out the reason for the inconsistency, let us take a look at scenario A in table 1. In this scenario, the overall average score, as well as the average scores of girls and boys in school 1 are all higher than those in school 2. A closer examination shows that the proportions of girls are different in scenarios A and B. In scenario A, 48% of students are girls in both schools. In scenario B, 40% and 60% of students are girls in the two schools, respectively. Is the difference in the proportions of girls sufficient to create this inconsistency? The answer is negative. In scenario C in table 1, although the proportions of girls are different in the two schools, the overall average scores and the average scores by gender are higher in school 1.

Examples in table 1 indicate that the results between overall analysis and subgroup analysis may be very different. Now we show what overall analysis and subgroup analysis actually mean.[1] Suppose we are interested in the treatment effect of a new drug D. We recruit some subjects and randomise them to two treatment groups (T and C). Subjects in groups T and C were administered drug D and placebo, respectively. After collecting the data, we calculate the average response of these two groups and use appropriate statistical methods (eg, two-sample t-test, Pearson's χ^2 test, and so on) to compare them. This is called the overall analysis. However, we suspect the response of a subject may depend on his/her age. We divide the subjects in the study population into several subgroups based on their ages and study the treatment effect within each age group. This kind of analysis is called the subgroup analysis. The subgroup analysis may offer us more information on the treatment effect of the new drug within each specific age group.

Results in table 1 indicate that even if in each subgroup, the new drug turns out to be better than the placebo, the overall response in the placebo group may be better than the new drug group. In this paper, we studied the reason for this counterintuitive phenomenon. The paper is organised as follows. Section 2 defines some notations. Section 3 gives a very general sufficient condition of consistency. We give some practical guidance in dealing with inconsistency in real studies in section 4.

NOTATIONS

We used the example in table 1 to develop our notation. However, our results apply to both continuous and categorical outcomes. Let Y_i denote the score of a randomly selected student from school i, and p_i denote

Table 1 Average scores of a maths exam in two schools

Scenario	School	Girls		Boys		Overall average score
		N	Average score	N	Average score	
A	1	48	91	52	81	85.8
	2	48	90	52	80	84.8
B	1	40	91	60	81	85.0
	2	60	90	40	80	86.0
C	1	48	91	52	81	85.8
	2	52	90	48	80	85.2

the proportion of girls in school i, i=1, 2. We define the following quantities: a_{10}=average score of girls in school 1, a_{11}=average score of boys in school 1, a_{20}=average score of girls in school 2, a_{21}=average score of boys in school 2.

Then the overall average scores of these two schools are $a_1=p_1a_{10}+(1-p_1)a_{11}$, and $a_2=p_2a_{20}+(1-p_2)a_{21}$, respectively. They are the weighted averages of the subgroup averages.

We also define some differences in the score:

(1) The differences between girls and boys within each school: $d_1=a_{10}-a_{11}$, $d_2=a_{20}-a_{21}$.

(2) The difference in girls (boys) between the two schools (subgroup differences):

$\Delta_0=a_{10}-a_{20}$,

$\Delta_1=a_{11}-a_{21}$.

(3) The overall difference between the two schools:

$\Delta=a_1-a_2$.

It is easy to prove that

$\Delta_0-\Delta_1=d_1-d_2$.

SUFFICIENT CONDITION OF CONSISTENCY

The inconsistency between the overall difference and the subgroup differences happens when Δ_0 and Δ_1 have the same sign but Δ has the opposite sign. In scenario B of table 1, the average scores of girls and boys in school 1 are higher than those in school 2. However, the overall average score of school 2 is higher. The inconsistency between the overall analysis and subgroup analysis happens in this case. Since each of them can be >0, =0 or <0, there are 27 possible combinations of the signs of Δ_0, Δ_1 and Δ. For the sake of completeness, we list all 27 combinations in table 2.

There are many redundancies in table 2. If we exchange the labels of those two schools, combinations 1–13 become combinations 15–27. Therefore, we do not need to consider combinations 15–27. There are still some other redundancies in combinations 1–14. For example, if we relabel girls and boys, combinations 4–6 become combinations 10–12. Combination 14 is of no interest in practice and will not be discussed further. Hence, we only consider combinations 1–9 and 13 in our discussion of (in)consistency.

The inconsistency between Δ_0, Δ_1 and Δ occurs if and only if one of the following occurs:

$\Delta_0>0$ and $\Delta_1>0$ but $\Delta<0$,

$\Delta_0>0$ and $\Delta_1>0$ but $\Delta=0$,

$\Delta_0>0$ and $\Delta_1=0$ but $\Delta<0$,

$\Delta_0=0$ and $\Delta_1=0$ but $\Delta>0$.

Combinations 2, 3, 5, 6 and 13 in table 2 satisfy the conditions above.

From the previous section, we can see that

$\Delta=d_1(p_1-p_2)+\Delta_1(1-p_2)+\Delta_0p_2=d_2(p_1-p_2)+\Delta_1(1-p_1)+\Delta_0p_1$.

Consider the following four cases:

(1) Two schools have the same proportion of girls, that is, $p_1=p_2$. The marginal difference is $\Delta=\Delta_0(1-p_1)+\Delta_1p_1=\Delta_0(1-p_2)+\Delta_1p_2$, which is a weighted average of the subgroup differences. There is no inconsistency between the subgroup and overall analyses. Scenario A in table 1 is an example of this case.

(2) There is no difference between the average scores of girls and boys in school 1, that is, d_1=0. The marginal difference is $\Delta=\Delta_0(1-p_2)+\Delta_1p_2$, which is a weighted average of the conditional differences. There is no inconsistency between the subgroup and overall analyses.

(3) There is no difference between the average scores of girls and boys in school 2, that is, d_2=0. The marginal difference is $\Delta=\Delta_0(1-p_1)+\Delta_1p_1$, which is a weighted average

Table 2 All possible combinations of signs of Δ_0, Δ_1 and Δ

Combination	Δ_0	Δ_1	Δ	Combination	Δ_0	Δ_1	Δ
1	>0	>0	>0	15	<0	<0	<0
2	>0	>0	=0	16	<0	<0	=0
3	>0	>0	<0	17	<0	<0	>0
4	>0	=0	>0	18	<0	=0	<0
5	>0	=0	=0	19	<0	=0	=0
6	>0	=0	<0	20	<0	=0	>0
7	>0	<0	>0	21	<0	>0	<0
8	>0	<0	=0	22	<0	>0	=0
9	>0	<0	<0	23	<0	>0	>0
10	=0	>0	>0	24	=0	<0	<0
11	=0	>0	=0	25	=0	<0	=0
12	=0	>0	<0	26	=0	<0	>0
13	=0	=0	>0	27	=0	=0	<0
14	=0	=0	=0				

Table 3 Numerical examples of Δ_0, Δ_1 and Δ when $p_1 \neq p_2$ and $d_1 d_2 \neq 0$

Case	p_1	p_2	d_1	d_2	Δ_0	Δ_1	Δ
1	0.4	0.5	2	2	1	1	1.2
2	0.4	0.5	−10	−10	1	1	0
3	0.4	0.5	−20	−20	1	1	−1
4	0.4	0.5	3	2	1	0	0.8
5	0.4	0.5	−5	-6	1	0	−1.4
6	0.4	0.5	−19	−20	1	0	−1.4
7	0.4	0.5	3.5	2	1	−0.5	0.6
8	0.4	0.3	1	−2	1	−2	0
9	0.4	0.5	−18	−20	1	−1	−1.8
13	0.4	0.5	20	20	0	0	2

of the subgroup differences. There is no inconsistency between the subgroup and overall analyses.

(4) Two schools have different proportions of girls, and the average scores of girls and boys are different within each school, that is, $p_1 \neq p_2$, $d_1 \neq 0$, $d_2 \neq 0$. This is the most general case in practice.

The first three cases indicate that $p_1 = p_2$ or $d_1 d_2 = 0$ is a sufficient condition of consistency between subgroup and overall analyses, as the overall difference is a convex combination of subgroup differences.

In table 3, we use numerical examples to show that if $p_1 \neq p_2$ and $d_1 d_2 \neq 0$, all combinations of 1–9 and 13 in table 2 may occur.

The following theorem gives a more general sufficient condition of consistency than the first three cases discussed above.

Theorem: given Δ_0 and Δ_1, for any p_1 and p_2 between 0 and 1, there always exists a p between 0 and 1 such that $\Delta = \Delta_0 p + \Delta_1 (1-p)$ if and only if $p_1 = p_2$ or $d_1 d_2 \leq 0$.

The proof of this theorem is available on request. Note that $d_1 d_2 = 0$ implies $d_1 d_2 \leq 0$.

Unfortunately, if we are only given the information that $p_1 \neq p_2$ and $d_1 d_2 > 0$, we cannot determine whether the inconsistency will happen. For example, combinations 1 and 3 satisfy the condition of $p_1 \neq p_2$ and $d_1 d_2 > 0$. In combination 1, the overall difference is consistent with the subgroup differences, while it is not in combination 3.

CONCLUSION AND DISCUSSION

Many publications of medical studies report the results of primary analysis based on all data and of subgroup analyses (with the same outcome in the primary analysis) based on partial data in the same study.[2] Sometimes, the result from the primary analysis may be inconsistent with the subgroup analysis. In this paper, we give a general sufficient condition of consistency between the overall and subgroup analyses. However, examples in table 3 indicate that it is impossible to give a general necessary condition of consistency. We need to check the consistency case by case.

Like the well-known Simpson's paradox,[3–6] the inconsistency between the overall and subgroup analyses seems to be counterintuitive for many people at first sight. Statistically speaking, the overall analysis and subgroup analysis use different parts of the data in the sample. The subgroup analysis uses only a partial sample of the study population, like the subgroup of girls in section 1. If the subgroup is not representative of the whole sample, inconsistency may occur. Both overall and subgroup analyses are valid methods to analyse data. They reveal different aspects of the data. The inconsistency is natural. We should interpret the results separately. It does not make sense to compare the results of the overall and subgroup analyses as they use different data. We can always write the overall analysis and subgroup analysis in the form of conditional expectations.[7] However, their conditional parts are different, and the results may be different.

From the example in section 1, we know that if two schools have the same proportions of girls, inconsistency will not happen. Like a randomised clinical trial, if the students were perfectly randomised to two schools, the proportion of girls will be very similar and the inconsistency will not happen in most cases. However, in most clinical trials, randomisation may not be balanced and inconsistency may still happen. On the other hand, if the data are from an observational study,[8] inconsistency may happen with high probability.

Another related topic is covariate adjustment in data analysis. For example, suppose the subjects were randomised into two treatment groups (active treatment and control). The primary outcome is binary (success or failure). We can use Pearson's χ^2 test to check the difference in success rates between the two groups. The odds ratio (OR) can be used to characterise the association between the treatment and the outcome. If we also have some other covariates, such as the age or gender of the subjects, we may also run a logistic regression using the treatment indicator and other variables as covariates. They are both valid methods to analyse the data.[9] However, the new OR may be different from the old one.

Contributors HW, CF and XMT: theoretical derivation; BW: manuscript drafting.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iDs
Bokai Wang http://orcid.org/0000-0002-2998-693X
Changyong Feng http://orcid.org/0000-0002-4432-1565

REFERENCES

1 Senn S. Conditional and marginal models: another view: Comment. *Statistical Science* 2004;19:228–30. doi:10.1214/088342304000000305
2 Wang R, Lagakos SW, Ware JH, *et al*. Statistics in medicine--reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007;357:2189–94.
3 Wang B, Wu P, Kwan B, *et al*. Simpson's paradox: examples. *Shanghai Arch Psychiatry* 2018;30:139–43.
4 Simpson EH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B* 1951;13:238–41. doi:10.1111/j.2517-6161.1951.tb00088.x
5 Yule GU. Notes on the theory of association of attributes in statistics. *Biometrika* 1903;2:121–34.
6 Heydtmann M. The nature of truth: Simpson's paradox and the limits of statistical data. *QJM* 2002;95:247–9.
7 Feller W. *An introduction to probability theory and its applications*. II. New York: Wiley, 1966.
8 Rosenbaum PR. *Observational studies*. 2nd ed. New York: Springer, 2002.
9 Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. 2nd ed. Hoboken, NJ: Wiley, 2002.

Hongyue Wang obtained her BS in Scientific English from the University of Science and Technology of China (USTC) in 1995, and a PhD in Statistics from the University of Rochester, USA, in 2007. She is currently a Research Associate Professor in the Department of Biostatistics and Computational Biology at the University of Rochester Medical Center in USA. Her main research interests include longitudinal data analysis, missing data, survival data analysis, and design and analysis of clinical trials. She has extensive and successful collaboration with investigators from various areas, including Infectious Disease, Nephrology, Neonatology, Cardiology, Neurodevelopmental and Behavioral Science, Radiation Oncology, Pediatric Surgery, and Dentistry. She has published more than 90 statistical methodology and collaborative research papers in peer-reviewed journals.