SOCIOLOGICAL EDUCATION

UDC 303.447.22

SERHII DEMBITSKII,

Candidate of Sciences in Sociology, Junior Research Fellow, Department of Sociological Methodology and Methods, Institute of Sociology, NAS of Ukraine

Heterogeneity assessment in meta-analysis (after the example of cross-national studies' data)

Аннотация

The article is dedicated to the principles of analysis of heterogeneity in the framework of meta-analysis. The author describes the essence of the heterogeneity estimation as well as considers the following problems: statistical testing with the help of Q-test; estimation of true dispersion of the values of effects with the help of T^2; estimation of proportions of the observed variability which points to the true variability of the values of effects with the help of I^2; calculation of confidential and predictional intervals. Based on the practical example with the use of relation of chances as the values of effects the author demonstrates the calculation of key indices of heterogeneity (results of the fourth wave of the European Social Survey were used as the empirical basis).

Keywords: *meta-analysis of the values of the effect, model of random effects, heterogeneity indices*

Introduction

In my latest publication, I've reviewed key notions and calculation basics in meta-analysis [Dembitskii, 2012]. At that special attention was given to two models of meta-analysis – to fixed and random effects models. In the scope of random effects model an important task is heterogeneity assessment, that allows: 1) to determine the share of variability in the effects values caused by differences of general totality out of which they were extracted; 2) to make a prognosis about the range of interval, under which fall most of the effects values, irrespective of the type of general totality they were extracted from; 3) to formulate variability factors hypotheses, that is to say of those relevant factors by which the general totality is differentiated. The latter opens an opportunity for a separate analysis of subgroups of contrasting effects values and, respectively, makes studies' output more flexible at the expense of binding to contextual features, which are important from the standpoint of the analyzed interconnection.

It should be noted that, that the following question will be of heterogeneity assessment, not its verification. Often does the agreement on the heterogeneity/homogeneity of results get accepted in a result of a statistic assessment (statistic method). Although in the scope of the article I am departing from Michael Borenstein's and his co-authors position [Borenstein et al., 2009], that connect the agreement on existence of heterogeneity of results or its absence with nature of the analyzed data (conceptual method). In the use of statistic method, it is conventional to begin with application of the model of fixed effects and then to check the results for heterogeneity, and in the case of their affirmation, to switch to random effects model. While in the conceptual method the researcher chooses one of the two models, uses arguments about the origins of the empiric material(it should be noted that effects values are extracted from different general totalities) to decide which one of the two models to pick. And only in the case of the random effects model being chosen does the heterogeneity assessment get conducted, as it is initially supposed to[1].

Heterogeneity assessment grounds

When it comes to heterogeneity assessment in the scope of meta-analysis, intrinsic variability of effects values is implied (on the contrary to random errors). At the same time, the observed variability includes in itself both components – random errors and intrinsic variability. The mechanism that allows to separate these components looks like this: 1) to calculate the total value of variability, which is observed between studies, which were handpicked for analysis; 2) to assess the expected variability in case of all the effects values residing only in a single general totality(that is in the case when the aggregated variability is based solely on random errors); 3) to find the difference between general and expected variability (the so-called excessive variation), that point to the heterogeneity or intrinsic variability value [Borenstein et al., 2009: p. 108].

Statistical review with the help of Q-test. Traditionally heterogeneity assessment begins with Q-statistic calculation, showing if the observed variability statistically differs from the expected, that is based on random errors.

Q — indicator, that has χ^2 allocation, with (*df*) degrees of freedom, equal to $k-1$, where *k* is equal to the number of effect variables [Corcoran, Littel, 2010: p. 303]. *Q* is a standardized, so it doesn't depend on the metric of the used effect variables. The basic formula of calculation of *Q* has the following form:

$$1.1.\, Q = \sum_{i=1}^{k} W_i \,(Y_i - M)^2 \,.$$

[1] For more information on selection of a meta-analysis model in the context of results heterogeneity assessment see: [Hunter, Schmidt, 2004: p. 393–399].

Equivalent formula is more suitable for calculations:

$$1.2.\, Q = \sum_{i=1}^{k} W_i Y_i^{\,2} - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i}.$$

The expected Q value for the case of all effects values belong to a single general totality, doesn't differ from df, because it is standardized. Therefore, the intrinsic variation in effects values can be calculated by subtracting from the expected Q its expected value, or df. If this value is negative or slightly exceeds 0 (if the magnitude of difference between observed and expected values, gets defined by the standard for the usage of χ^2 criteria method — with the help of p-value), then the review didn't show any statistically significant results. Otherwise, the heterogeneity decision grants the opportunity for the further analysis.

Q-test is a null hypothesis review (according to which all studies describe all the same effect or belong to a single general totality), that is connected to the errors that are also typical for other tests of statistical significance. First, if the statistically significant results point out at heterogeneity, then insignificant do not yet prove its absence — possible explanation may be hiding in the small sample size or the big result dispersions of separate studies. Second, Q-test results are exclusively used for null hypothesis review and, therefore, aren't indicators of intrinsic variability value. This task is solved by other indices, which will be discussed below.

Effects values intrinsic dispersion assessment with the help of T^2. T^2 index brings back the variability assessment from standardized scale to original metric of effect values:

$$1.3.\, T^2 = \frac{Q - df}{C},$$

where $C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}.$

Intrinsic dispersion value of the effects values can't be lower than 0, while T^2 can assume a negative value (at $Q < df$), which is connected with the errors of sampling. In this case T^2 assumes the value of 0. If $Q > df$, then T^2 will be a positive value, which is based on two factors. First of the two — excessive variability value ($Q - df$), second — dimension of the scale where effects values are measured.

To assess effects values standard deviation T is used, which can be found with a simple square rooting of the T^2. In finding of T^2 assumptions about the form of distribution of the effects values are not needed, but if there are grounds to consider it normal, then T may be used for interval location, in which falls the specified proportion of all the effects values. For example, 95% of all possible values falls in the interval that is equal to $1{,}96 \times T$.

With that, this method of constructing of effects values distribution is justified only in the case of these values and T being assessed correctly. So, in practice, the construction of the corresponding intervals for distribution of effects values, errors made in the assessment of both parameters must be taken into account. Grounds of the corresponding calculations are reviewed below.

Observed variability proportion assessment, which points out at intrinsic variability of the effects values using I^2. I^2 is an index derived from Q:

$$1.4.\ I^2 = \left(\frac{Q-df}{Q}\right)\times 100\%.$$

As well as Q, I^2 index is standardized, thus it doesn't depend on the original scale of effects values and varies between 0% and 100%. An important distinctive feature of this statistic is also the fact that it isn't dependent on the quantity of the analyzed effects values [Borenstein et al., 2009: p. 109–119].

It should be remembered that — despite I^2 allowing assessment of the intrinsic variability proportion, it doesn't give out any other information. Therefore, if the correspondent value is close to 100%, it does not yet tell anything about the effects values distribution, which in this case may fall in a narrow as well as a wide interval. I^2 interpretation guidelines, listed in the work of Higgins [Higgins et al., 2003: p. 559][1], should be also reviewed in the context of the size of *proportion* of intrinsic variability assessment, not its absolute value.

Heterogeneity indices comparison. All of the above listed heterogeneity indices are based on Q (in the relation to df). With that, they are designed for solution of various problems, which makes usage of each of them — justified and necessary. The following is their comparative statistics (see. table 1).

Confidence and predictive intervals

Confidence intervals. So far as T^2 and I^2 are estimates of the corresponding parameters, and not intrinsic indices, it makes sense to calculate confidence intervals for variability of the effects values interpretation, in which with an estimated probability must fall the intrinsic variability and its proportion.

Table 1

Heterogeneity indices

Index and its values span	Sample size dependency	Scale dependency	Solved question
$0 \leq Q$	+	–	Can the observed variability be explained exclusively by random errors?
$0 \leq T^2$	–	+	What is the value of intrinsic variability of effects values?
$0 \leq T$	–	+	Which interval do the most effects values fall in?
$0\% \leq I < 100\%$	–	–	What proportion of the general variability is intrinsic, and therefore not stipulated by random errors?

[1] 0% — heterogeneity is absent, 25% — low heterogeneity, 50% — average heterogeneity, 75% — high heterogeneity.

If to consider effects values to be distributed normally (suitable for most sample groups), a standard T^2 error can be assessed this way.

First calculate the intermediate value A:

$$2.1.\ A=\left[df+2\left(sw1-\frac{sw2}{sw1}\right)T^2+\left(sw2-2\left(\frac{sw3}{sw1}\right)+\frac{(sw2)^2}{(sw1)^2}\right)T^4\right],$$

where

$$sw1=\sum_{i=1}^{k}W_i\,,$$

$$sw2=\sum_{i=1}^{k}W_i^2\,,$$

$$sw3=\sum_{i=1}^{k}W_i^3\,.$$

After what T^2 dispersion V_{T^2} can be found:

$$2.2.\ V_{T^2}=2\times\left(\frac{A}{C^2}\right).$$

The standard error (SE_{T^2}) will be correspondingly equal to $\sqrt{V_{T^2}}$.

But since the distribution T^2 doesn't correspond to the normal one good enough, calculation of the confidential intervals by multiplying the standard error by ±1,96 won't lead to an accurate enough result until a big enough sample groups are used. One of the easiest methods of solving this problem is the following:

If $Q>(df+1)$, find B according to the formula:

$$2.3.\ B=0{,}5\times\frac{\ln(Q)-\ln(df)}{\sqrt{2Q}-\sqrt{2\times df-1}}.$$

If $Q\le(df+1)$, using formula:

$$2.4.\ B=\sqrt{\frac{1}{2\times(df-1)\times\left(1-\left(\frac{1}{3\times(df-1)^2}\right)\right)}}.$$

Next, calculate the intermediate values L and U:

$$2.5.\ L=\exp\left(0{,}5\times\ln\left(\frac{Q}{df}\right)-1{,}96\times B\right)$$

and

$$2.6.\ U=\exp\left(0{,}5\times\ln\left(\frac{Q}{df}\right)+1{,}96\times B\right).$$

Finally, the confidence intervals for the intrinsic values of the dispersion effects can be found as follows:

$$2.7.\ LL_{T^2} = \frac{df \times (L^2 - 1)}{C}$$

and

$$2.8.\ UL_{T^2} = \frac{df \times (U^2 - 1)}{C}.$$

In finding the confidence intervals for the standard deviation of effects values, square root of the corresponding values of the variance estimation should be extracted:

$$2.9.\ LL_T = \sqrt{LL_{T^2}}\ ;$$

$$2.10.\ UL_T = \sqrt{UL_{T^2}}\ .$$

During the definition of the confidence intervals for I^2 conditions and calculation formulas of B, L and U are the same as in the case of T^2. If $Q > (df + 1)$, 2.3 formula is used. If $Q \leq (df + 1)$, 2.4 formula is used.

Discovery of the intermediate values is identical for formulas 2.5 and 2.6. Then a 95-percent confidence interval can be found:

$$2.11.\ LL_{I^2} = \left(\frac{L^2 - 1}{L^2}\right) \times 100\%$$

and

$$2.12.\ UL_{I^2} = \left(\frac{U^2 - 1}{U^2}\right) \times 100\%.$$

Any of the values for LL and UL, that are less than zero, should be equated to zero. If the lower limit of the interval is greater than zero, then I^2 should be statistically significant. Although, since I^2 is based on Q, while the sampling distribution of Q is better explored one, than the sampling distribution of I^2, a more reliable method of statistical significance assessment of I^2 is exactly the Q-test.

Predictive intervals. Very often the main goal of meta-analysis is the calculation of a weighted average of effects values and its confidence intervals. And while it is an important task, solving it doesn't tell anything about the distribution of intrinsic effects values around the mean value. In the fixed effects model this method (finding only a mean value and its confidence intervals) is justified, because existence of a same intrinsic effect value for all studies is assumed. In turn, in the random effects model not only the intrinsic mean of effects values should be assessed, but their distribution around as well. The latter relates to definition of the predictive intervals.

When confidence intervals are being defined it is necessary to take into account errors made during assessment of the effects values and T, calculation of predictive intervals is carried out the following way:

$$2.13.\ LL_{pred} = M^* - t_{df}^{\alpha} \sqrt{T^2 + V_{M^*}}\ ,$$

$$2.14.\ UL_{pred} = M^* + t_{df}^{\alpha} \sqrt{T^2 + V_{M^*}}\ ,$$

where M^* — sample weighted mean of effects values, T^2 — assessment of the intrinsic effects values variance and V_{M^*} — M^* variance. Factor t^{α}_{df} — value from distribution of critical values of *t*-Student for the corresponding probability and amount of degrees of freedom (*df* = *k* — 1, where *k* — amount of effects values in the sample group).

Comparison of confidence and predictive intervals of the weighted mean. It its important to remember, that confidence and predictive intervals solve different tasks. The first type is designed to define the accuracy of assessment of the weighted mean of effects values, and the second — to define dispersion of all possible effects values around the sample mean.

A significant difference between the two types of intervals is showed in the way how they change with the sample group's increase. Thus, the confidence interval will tend to zero, while the predictive interval will decrease until a certain moment, after which the changes will practically stop. It is related to the predictive interval being based on, among other things, on T^2, a value that does not depend on size of the sample group [Borenstein et al., 2009: p. 122–133].

A practical example: respondent's gender and his/hers chances of employment abroad

For example of calculation of heterogeneity indices are used the same empirical data of the fourth wave of European social studies', as in the previous article [Дембицкий, 2012: c. 169–172] (see table 2). On the grounds of clarity in the table were included only those indices that are directly used in the current calculations.

I will also remind that, the weighted median magnitude if effects values for the odds ration (from this point onward — M^*) equals to 0,53, and it's variance — 0,004. After the implementation of the inverse transformation (both indices are calculated after a logarithmic transformation) the mean will be equal to 1,7, and the variance — 1,0.

First, with the help of *Q*-test you need to check the null hypothesis that suggests that all the variability between the sample group's effects values is caused by random errors. For this formula 1.2 is used:

$$Q = 244{,}01 - \frac{331{,}58^2}{639{,}07} = 71{,}97.$$

With the degrees of freedom being equal to 29, and level α, equal to 0,05, the critical value of distribution χ^2 is equal to 42,6. The latters allows to make a conclusion about the heterogeneity of the effects values.

Let's make an assessment of the intrinsic effects values variance (formula 1.3):

$$df = 30 - 1 = 29,$$

$$C = 639{,}07 - \frac{15526{,}32}{639{,}07} = 614{,}77,$$

$$T^2 = \frac{71{,}97 - 29}{614{,}77} = 0{,}07.$$

Table 2

Data, necessary for effects values heterogeneity analysis

Country	W_i	W_i^2	W_i^3	$W_i Y_i$	$W_i Y_i^2$
Belgium	23,38	546,62	12780,08	10,10	4,36
Bulgaria	23,32	543,82	12681,94	4,83	1,00
Switzerland	26,92	724,69	19508,56	11,09	4,57
Cyprus	13,36	178,49	2384,62	9,26	6,42
Czech Republic	27,94	780,64	21811,18	7,12	1,82
Germany	19,05	362,90	6913,29	14,57	11,15
Denmark	16,60	275,56	4574,30	3,02	0,55
Estonia	25,27	638,57	16136,74	31,94	40,37
Spain	33,47	1120,24	37494,46	0,00	0,00
Finland	20,04	401,60	8048,10	10,04	5,03
France	21,51	462,68	9952,25	13,12	8,00
Great Britain	27,79	772,28	21461,78	14,90	7,98
Greece	18,64	347,45	6476,46	4,75	1,21
Croatia	16,51	272,58	4500,30	12,56	9,56
Hungary	15,17	230,13	3491,06	9,74	6,25
Ireland	41,94	1758,96	73770,93	12,58	3,77
Israel	27,56	759,55	20933,30	0,28	0,00
Lithuania	26,12	682,25	17820,48	16,90	10,93
Latvia	25,22	636,05	16041,14	21,54	18,39
Netherlands	20,53	421,48	8653,00	16,55	13,34
Norway	3,59	12,89	46,27	1,18	0,39
Poland	20,58	423,54	8716,38	14,78	10,61
Portugal	22,33	498,63	11134,38	8,31	3,09
Romania	19,60	384,16	7529,54	6,02	1,85
Russia	7,62	58,06	442,45	12,73	21,25
Sweden	26,54	704,37	18694,02	13,62	6,98
Slovenia	8,80	77,44	681,47	6,27	4,47
Slovakia	29,50	870,25	25672,38	28,85	28,22
Turkey	7,17	51,41	368,60	–1,69	0,40
Ukraine	23,00	529,00	12167,00	16,63	12,02
Σ	**639,07**	**15526,32**	**410886,45**	**331,58**	**244,01**

The corresponding standard deviation will be equal to:

$$T = \sqrt{0{,}07} = 0{,}26.$$

Now let's calculate the confidence interval for T^2. In the assumed case of normality of effects values variance first *sw1, sw2* and *sw3* have to be found:

$$sw1 = 639{,}07$$

$$sw2 = 15526{,}32$$

$$sw3 = 410886{,}45$$

After what A can be calculated (formula 2.1), V_{T^2} (formula 2.2) andSE_{T^2} :

$$A = 29 + \left(639{,}07 - \frac{15526{,}32}{639{,}07}\right)0{,}07 +$$

$$+\left(15526{,}32 - 2\left(\frac{410886{,}45}{639{,}07}\right) + \frac{15526{,}32^2}{639{,}07^2}\right)0{,}07^2 = 187{,}40,$$

$$V_{T^2} = 2 \times \left(\frac{187{,}40}{614{,}77^2}\right) = 0{,}00099,$$

$$SE_{T^2} = \sqrt{0{,}00099} = 0{,}032.$$

Hereof, the 95-percent confidence interval will be equal:

$$C.i. = \pm\, 1{,}96 \times 0{,}032 = \pm\, 0{,}063.$$

If normality of the effects values variance isn't assumed then the calculation will take a different view. Since $Q = 71{,}97 > 30 = (df + 1)$, for the definition of B we use formula 2.3:

$$B = 0{,}5 \times \frac{\ln(71{,}97) - \ln(29)}{\sqrt{2 \times 71{,}97} - \sqrt{2 \times 29 - 1}} = 0{,}10.$$

The intermediate values (formulas 2.5 and 2.6 correspondingly):

$$L = \exp\left(0{,}5 \times \ln\left(\frac{71{,}97}{29}\right) - 1{,}96 \times 0{,}10\right) = 1{,}29,$$

$$U = \exp\left(0{,}5 \times \ln\left(\frac{71{,}97}{29}\right) + 1{,}96 \times 0{,}10\right) = 1{,}92.$$

Therefore, limits of the 95-percent confidence interval are equal to (according to formulas 2.7 and 2.8 correspondingly):

$$LL_{T^2} = \frac{29 \times (1{,}29^2 - 1)}{614{,}77} = 0{,}03,$$

$$UL_{T^2} = \frac{29 \times (1{,}92^2 - 1)}{614{,}77} = 0{,}13.$$

If to go from metrics of the natural logarithm to original metrics of the odds ratio, then variance (T^2) will be equal to 1,07, lower interval — 1,03, upper interval — 1,14.

On the penultimate stage it will be revealed what proportion of the general variability is genuine (formula 1.4):

$$I^2 = \left(\frac{71{,}97 - 29}{71{,}97}\right) \times 100\% = 59{,}7\%.$$

While calculating the confidence intervals for I^2, values for L and U will be the same as in the calculation of confidence intervals for T^2. Therefore an immediate transition to the calculation of the interval's limits is permitted (formulas 2.11 and 2.12 correspondingly):

$$LL_{I^2} = \left(\frac{1{,}29^2-1}{1{,}29^2}\right)\times 100\% = 39{,}9\%,$$

$$UL_{I^2} = \left(\frac{1{,}92^2-1}{1{,}92^2}\right)\times 100\% = 73{,}0\%.$$

Finally, predictive intervals for the effects values are to be calculated (formulas 2.13 and 2.14 correspondingly), considering $M^* = 0{,}53$, $V_{M^*} = 0{,}04$, and $t_{28}^{0,05} = 1{,}7$ (degrees of freedom equal to 28, α level — 0,05).

$$LL_{pred} = 0{,}53 - 1{,}7\sqrt{0{,}070{,}004} = 0{,}07,$$

$$UL_{pred} = 0{,}53 + 1{,}7\sqrt{0{,}070{,}004} = 0{,}99.$$

Respectively, after the transformation into a scale of odds ratio the lower limit will be equal to 1,07, upper — 2,69.

Now let's review all the data combined (see table 3). As it was stated earlier on, Q value (71,97) allows to make a conclusion about heterogeneity of the results with the error probability not exceeding 5%. The magnitude of the intrinsic variability in the original scale of effects values is approximately 1, which makes about 60% of the total variability. The latter value isn't accurate enough and with a probability high enough falls in the interval between 40% and 73%.

Table 3

The resulting data describing the effects values heterogeneity *

Index	Value	Confidence/predictive interval	
		Lower limit	Upper limit
M^*	0,53(1,7)	0,41(1,5)	0,65(1,9)
Q	71,97	–	–
T^2	0,07(1,07)	0,03(1,03)	0,13(1,14)
T	0,26(1,3)	0,17(1,19)	0,36(1,43)
I	59,7%	39,9%	73,0%
PI^{**} *for* M^*	–	0,07(1,07)	0,99(2,69)

* For the cases when the index is calculated in the logarithmic scale, original metric effects values are shown in brackets (odds ratio).

** Predictive interval.

Despite, the resulting mean being equal to 1,7, that says about the weak connection between gender and odds of employment abroad, 95% of all the possible effects values are distributed in the interval from 1,07 (almost full absence of connection) to 2,69 (average connection).

Summary

Heterogeneity analysis is an important part of the accurate interpretation of the meta-analysis results. And in the case of cross-national and cross-cultural studies variability research becomes altogether vital part of this research method, allowing to assess and describe differences between units of analysis in strict terms. However assessment of the intrinsic variability, its proportions among the aggregate variability, definition of the predictive interval and calculation if other indices aren't the final point of the heterogeneity analysis.

After the results' heterogeneity has been established and its indices calculated,the above discussed ones, there are two ways of the further analysis. First is to divide the effects values into homogeneity subgroups and their analysis, second — to use metaregression [Leeuw, Hox, 2003: p. 336–339]. Both these instruments will be reviewed in one of the nearest issues of the magazine.

Sources

Dembitskii S. Metaanaliz: kliuchevye poniatiya i osnovy vychislenii (na primere dannykh kross-natsional'nykh issledovanii) / S. Dembitskii// Sotsiologiia: teoriia, metody, marketyng. — 2012. — № 3. — S. 160–174.

Borenstein M. Introduction to Meta-Analysis / M. Borenstein, L. Hedges, J. Higgins, H. Rothstein. — N. J. : Wiley, 2009.

Corcoran J. Meta-Analyses / J. Corcoran, J. Littel // The Handbook of Social Work Research Methods / ed. by B. Thyer. — Los Angeles ; London ; New Delhi ; Singapore ; Washington (DC) : SAGE, 2010. — P. 299–312.

Higgins J. Measuring inconsistency in meta-analyses / J. Higgins, S. Thompson, J. Deeks, D. Altman // British Medical Journal. — 2003. — Vol. 327. — P. 557–560.

Hunter J. Methods of meta-analysis: Correcting error and bias in research findings / J. Hunter, F. Schmidt. — Thousand Oaks ; London ; New Delhi : SAGE, 2004. — 582 p.

Leeuw E. The Use of Meta-Analysis in Cross-National Studies / E. Leeuw, J. Hox // Cross-Cultural Survey Methods / ed. by J. Harkness, F. Van de Vijver, P. Mohler. — N. J. : Wiley, 2003. — P. 329–345.